

Adapting the Cube algorithm for balanced random imputation in surveys

Guillaume Chauvet · Jean-Claude Deville · David Haziza

Received: date / Accepted: date

Abstract Random imputation methods are often used in practice because they tend to preserve the distribution of the variable being imputed. As a result, estimators of quantiles are asymptotically unbiased. One drawback of random imputation methods is that they introduce an additional amount of variability, called the imputation variance, due to the random selection of residuals. To overcome the problem, Chauvet et al (2010) proposed a random balanced imputation method, which eliminates the imputation variance while preserving the distribution of the variable being imputed. In this paper, we describe an algorithm designed for selecting random residuals under constraints, which is an adaptation of the Cube method proposed by Deville and Tillé (2004) in the context of balanced sampling.

Keywords Balanced sampling · Imputation variance · Random imputation · Stratified balanced sampling

Guillaume Chauvet
Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France
Tel.: +0033-299 053 323
Fax: +0033-299 053 205
E-mail: chauvet@ensai.fr

Jean-Claude Deville
Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France
Tel.: +0033-299 053 314
Fax: +0033-299 053 205
E-mail: deville@ensai.fr

David Haziza
Département de mathématiques et de statistique, Université de Montréal, Montréal, Québec, H3C 3J7, Canada
Tel.: +1-514 343 6705
Fax: +1-514 343 5700
E-mail: David.Haziza@umontreal.ca

1 Introduction

Imputation is the main technique for treating item nonresponse in surveys. It consists of replacing missing values with artificial values in order to reduce, as much as possible, the bias and the variance due to nonresponse. In order to preserve the distribution of the variable being imputed, it is customary to use some form of random imputation, in which case, estimators of population totals (or means) and estimators of quantiles (e.g., a median) are asymptotically unbiased. However, unlike deterministic imputation methods, random methods lead to potentially inefficient estimators because this type of methods introduces an additional amount of variability (called the imputation variance) due to the random selection of residuals. Chauvet et al (2010) proposed the use of balanced random imputation methods, which considerably reduce (or eliminate) the imputation variance while preserving the distribution of the variable being imputed. Balanced random imputation consists of selecting randomly residuals, while satisfying appropriate constraints. It can be applied to any type of random imputation method (e.g., random regression imputation) under any type of sampling design and can be used to impute continuous or categorical variables. In this paper, we show that the selection of residuals under constraints can be performed by adapting the Cube algorithm originally proposed by Deville and Tillé (2004) in the context of balanced sampling; see also Chauvet and Tillé (2006, 2007) and Chauvet (2009).

The outline of the paper is as follows: in section 2, we introduce some notation and present the imputed estimator of a total. Random imputation is discussed in section 3. It is shown in section 4 that balanced sampling techniques may be used for the selection of the residuals so that the imputation variance is virtually eliminated. The selection algorithm is described in details in section 5. Finally, in section, we conclude with some remarks.

2 Notation

Let U be a finite population consisting of N elements. We consider the problem of estimating a population total $Y = \sum_{i \in U} y_i$, where y_i denotes the i -th value of the variable of interest y , $i = 1, \dots, N$. We select a sample s , of size n , according to a given sampling design $p(s)$. Let π_i denote the first-order inclusion probability of unit i in the sample and let $w_i = 1/\pi_i$ denote its design weight. In the absence of nonresponse, a design-unbiased estimator of Y is the expansion estimator given by

$$\hat{Y}_\pi = \sum_{i \in s} w_i y_i. \quad (1)$$

In the presence of nonresponse to item y , we observe the y -values for a subset of the sampled units only. Let y_i^* denote the imputed value used to replace the

missing y_i . We define an imputed estimator \hat{Y}_I as

$$\hat{Y}_I = \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) y_i^*, \quad (2)$$

where r_i is a response indicator attached to unit i such that $r_i = 1$ if unit i responds to item y and $r_i = 0$, otherwise. Also, let s_r be the random set of respondents of size n_r and s_m the random set of non-respondents of size n_m .

3 Random imputation

Most of the imputation methods used in practice can be motivated by the so-called imputation model

$$m : y_i = f(\mathbf{z}_i; \beta) + \sigma \sqrt{v_i} \epsilon_i, \quad (3)$$

where $f(\cdot)$ is a given function, $\mathbf{z} = (z_1, \dots, z_K)'$ is a K -vector of auxiliary variables available at the imputation stage for all the sampled units, β is a vector of unknown parameters, σ^2 is an unknown parameter and v_i is a known constant. The ϵ_i 's are independent and identically distributed random variables with mean 0 and variance 1.

Based on (3), the imputed values y_i^* under random imputation are given by

$$y_i^* = f(\mathbf{z}_i; \hat{\beta}_r) + \hat{\sigma} \sqrt{v_i} \epsilon_i^*, \text{ for } i \in s_m, \quad (4)$$

where $\hat{\beta}_r$ is an estimator of β based on the responding units and $\hat{\sigma}$ is an estimator of σ . It is natural to select (usually with replacement) the random component ϵ_i^* from the set, $E_j = \{\tilde{e}_j; j \in s_r\}$, of standardized residuals observed from the responding units, with probabilities

$$\Pr(\epsilon_i^* = \tilde{e}_j) = \omega_j / \sum_{l \in s} \omega_l r_l, \quad (5)$$

where $\tilde{e}_j = e_j - \bar{e}_r$, $e_j = \frac{1}{\hat{\sigma} \sqrt{v_j}} (y_j - f(\mathbf{z}_j; \hat{\beta}_r))$, $\bar{e}_r = \sum_{j \in s} \omega_j r_j e_j / \sum_{j \in s} \omega_j r_j$ and ω_j is an imputation weight attached to unit j . Several choices of ω_j are possible: for example, the choice $\omega_j = w_j$ leads to the customary survey weighted random imputation, whereas the choice $\omega_j = 1$ leads to unweighted random imputation. Deterministic imputation under model (3) is obtained from (4) by setting $\epsilon_i^* = 0$ for all i .

Random regression imputation is obtained from (4) by setting $f(\mathbf{z}_i; \beta) = \mathbf{z}_i' \beta$. An important special case of random regression imputation in practice is random hot-deck imputation within classes. It consists of first partitioning the sample into K imputation classes, $s_1, \dots, s_k, \dots, s_K$. Within a class, a missing value is replaced by the value of a respondent selected randomly (with replacement) from the set of respondents within that class. Imputations are

performed independently across classes. Let $z_{ki} = 1$ if unit i belongs to class k and $z_{ki} = 0$ otherwise, with $k = 1, \dots, K$. Random hot-deck imputation within classes is obtained from (4) by setting $f(\mathbf{z}_i; \beta) = \mathbf{z}_i' \beta$ with $\mathbf{z}_i = (z_{1i}, \dots, z_{Ki})'$ and $v_i = v_k$ if i belongs to class k .

Using the imputed values (4) in (2) leads to

$$\begin{aligned} \hat{Y}_I &= \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) f(\mathbf{z}_i; \hat{\beta}_r) \\ &\quad + \sum_{i \in s} w_i (1 - r_i) \hat{\sigma} \sqrt{v_i} \epsilon_i^* \\ &= \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i (1 - r_i) f(\mathbf{z}_i; \hat{\beta}_r) \\ &\quad + \sum_{i \in s} w_i (1 - r_i) \hat{\sigma} \sqrt{v_i} \sum_{j \in s} r_j d_{ji} \tilde{e}_j, \end{aligned} \quad (6)$$

where $d_{ji} = 1$ if the residual \tilde{e}_j was selected for imputing the missing value y_i , and $d_{ji} = 0$ otherwise.

Let the subscripts p , q and I indicate the sampling mechanism, the non-response mechanism and the imputation mechanism, respectively. The total variance of \hat{Y}_I given in (2) can be expressed as

$$V(\hat{Y}_I) = V_p E_q E_I(\hat{Y}_I | s) + E_p V_q E_I(\hat{Y}_I | s) + E_p E_q V_I(\hat{Y}_I | s). \quad (7)$$

The first term on the right hand side of (7) is the sampling variance, the second term is the nonresponse variance, whereas the third term is the imputation variance. Note that the imputation variance is due to the third term in (2) only and is given by

$$E_p E_q V_I(\hat{Y}_I | s) = E_p E_q \left[\sum_{i \in s} w_i^2 (1 - r_i) v_i \frac{\sum_{i \in s} w_i r_i \tilde{e}_i^2}{\sum_{i \in s} w_i r_i} \right]. \quad (8)$$

From (8), we note that the magnitude of the imputation variance will be small if (i) the response rate is high, in which case the term $\sum_{i \in s} w_i^2 (1 - r_i) v_i$ is likely to be small (in fact, under full response, we have $r_i = 1$ for all i and so this term vanishes) or if (ii) the residuals \tilde{e}_i are small, which indicates that the imputation model fits the data well. Otherwise, the contribution of the imputation variance to the total variance may be appreciable.

For example, consider the case of simple linear regression imputation. Random simple linear regression imputation is obtained from (4) by setting $\mathbf{z}_i = (1, z_i)'$ and $v_i = 1$. Deterministic simple linear regression is further obtained by setting $\epsilon_i^* = 1$ for all i . Let $V_D(\hat{Y}_I)$ and $V_R(\hat{Y}_I)$ denote the total variance of \hat{Y}_I under deterministic and random simple linear regression imputation, respectively. Assume that the sample s is selected according to simple random sampling and that the nonresponse mechanism is uniform (that is, all the units have equal response probabilities, p say). Then, the additional

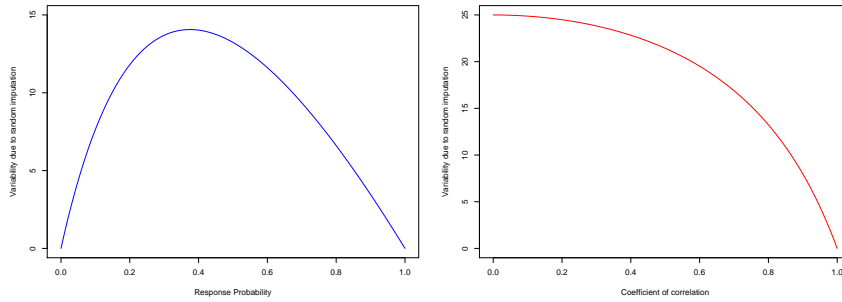


Fig. 1 Additional amount of variability (in %) due to random imputation. The left, blue curve corresponds to $\rho_{yz} = 0.8$, while the right, red curve corresponds to $p = 0.5$.

amount of variability (in %) due to random imputation, $C = \frac{V_R(\hat{Y}_I) - V_D(\hat{Y}_I)}{V_D(\hat{Y}_I)}$, can be approximated by

$$C \simeq \frac{p(1-p)(1-\rho_{yz}^2)}{1-(1-p)\rho_{yz}^2}, \quad (9)$$

provided the sample size n is sufficiently large, where ρ_{yz} denotes the coefficient of correlation between y and z . Figure 1 shows the additional amount of variability (in %) due to random imputation for a fixed value of ρ_{yz} ($= 0.8$) on the left-hand side, and for a fixed value of p ($= 0.5$) on the right-hand side. It is clear from Figure 1 that the contribution of the imputation variance is increasing in $[0, p_{max}]$, where $p_{max} = \sqrt{\frac{1-\rho_{yz}^2}{2-\rho_{yz}^2}}$ is the value for which C in (9) is maximum, and decreases in the interval $(p_{max}, 1]$. Note that when $\rho_{yz} = 0.8$ we have $p_{max} \simeq 0.51$. Also, it is clear that the contribution of the imputation variance decreases as the coefficient of correlation between y and z increases, as expected.

4 Balanced random imputation

In order to eliminate the imputation variance, Chauvet et al (2010) proposed a balanced random imputation method which consists of selecting the residuals ϵ_i^* so that the following equation is (approximately) satisfied:

$$\sum_{i \in s} w_i(1-r_i)\sqrt{v_i}\epsilon_i^* = 0, \quad (10)$$

which may be alternatively written as

$$\sum_{i \in s} w_i(1-r_i)\sqrt{v_i} \sum_{j \in s} r_j d_{ji} \tilde{\epsilon}_j = 0. \quad (11)$$

Table 1 Population of cells used for the random selection of residuals

	1	...	j	...	n_r
1	(ψ_{11}, \tilde{e}_1)	...	(ψ_{1j}, \tilde{e}_j)	...	$(\psi_{1n_r}, \tilde{e}_{n_r})$
...
i	(ψ_{i1}, \tilde{e}_i)	...	(ψ_{ij}, \tilde{e}_j)	...	$(\psi_{in_r}, \tilde{e}_{n_r})$
...
n_m	$(\psi_{n_m 1}, \tilde{e}_1)$...	$(\psi_{n_m j}, \tilde{e}_j)$...	$(\psi_{n_m n_r}, \tilde{e}_{n_r})$

If the equation (10) is exactly satisfied, then the imputation variance is completely eliminated and the resulting estimator is fully efficient; see Kim and Fuller (2004). In some situations, it is not possible to satisfy (10) exactly but only approximately. In this case, the imputation variance is not completely eliminated but it is expected to be significantly reduced. Additional constraints may be added for the selection of the residuals if it is desired to eliminate the imputation variance for other parameters, for example, the regression coefficient β in (3); see Chauvet et al (2010).

Equation (10) may be seen as a *balancing equation* which is imposed in the selection of the random residuals ϵ_i^* . That is, the random residuals ϵ_i^* must be selected with replacement from the set E_r of standardized residuals observed from the responding units, such that equation (10) holds. Consequently, we may think of the Cube Method proposed by Deville and Tillé (2004) for the selection of the random residuals. Unfortunately, the Cube Method was originally designed to select without replacement samples from a population, while respecting balancing constraints. The setting is different in case of balanced imputation, since with replacement of residuals is implied. That is, a same respondent may be selected several times to be used as a donor. We first show that with replacement balanced sampling may be alternatively seen as without replacement balanced sampling within a population of cells.

For the purpose of balanced random imputation, we consider the $n_m \times n_r$ population of cells given in Table 1, in which each cell (i, j) is given the value of the standardized residual \tilde{e}_j and the probability of selection $\psi_{ij} = \frac{\omega_j}{\sum_{l \in s} \omega_l r_l}$. Let U^* denote the $n_m \times n_r$ population of cells. A random imputation obtained from (4) may alternatively be seen as selecting a random sample s^* of cells in U^* without replacement, where the non-respondent i is given the residual associated to the respondent j if the cell (i, j) is selected in s^* . The sample must be drawn so that (i) exactly one cell per row is selected in s^* , since one residual exactly must be selected for each nonrespondent, and (ii) each cell has a probability of selection equal to ψ_{ij} , in line with (5).

The constraint (i) of selecting exactly one cell in row i^* may be written as

$$\sum_{j=1}^{n_r} d_{i^*j} = 1, \quad i^* = 1, \dots, n_m, \quad (12)$$

and can be alternatively written as a system of n_m balancing equations

$$\sum_{(i,j) \in s^*} \frac{\mathbf{x}_{ij}}{\psi_{ij}} = \sum_{(i,j) \in U^*} \mathbf{x}_{ij} \quad (13)$$

on a n_m vector of variables

$$\mathbf{x} = (x^1, \dots, x^{n_m})', \quad (14)$$

where the variable x^{i^*} takes the value $x_{ij}^{i^*} = \psi_{ij} \delta_{i^*i}$ on the cell (i, j) , and δ_{i^*i} equals 1 if $i^* = i$ and 0 otherwise. The equation (10) may also be written as the balancing equation

$$\sum_{(i,j) \in s^*} \frac{x_{ij}^0}{\psi_{ij}} = \sum_{(i,j) \in U^*} x_{ij}^0, \quad (15)$$

with $x_{ij}^0 = w_i \sqrt{v_i} \psi_{ij} \tilde{e}_j$ for the cell (i, j) , since it may be easily shown that $\sum_{(i,j) \in U^*} x_{ij}^0 = 0$. Selecting a sample s^* balanced on variables $\tilde{\mathbf{x}} = (\mathbf{x}', x^0)'$ with inclusion probabilities ψ_{ij} ensures that conditions (i) and (ii) as well as equation (10) are exactly satisfied and, as a result, that the variance imputation is eliminated.

An algorithm adapting the Cube method to the random selection of residuals as described above is given in section 5. In practice, note that there may exist no sample s^* such that both equations (13) and (15) are exactly satisfied. The Cube method then involves a rounding process called the landing phase in order to end the sampling while exactly respecting the inclusion probabilities. The landing phase may be performed by successively relaxing the balancing constraints, which has the advantage of permitting the selection of a sample in a reasonable amount of time, even if the number of balancing variables is large, see Tillé (2006, p. 163). A careful treatment of the landing phase is needed since the balancing equations (13) must be preserved until the end of the selection procedure. That is, exactly one respondent must be selected for each non-respondent. It is shown in section 5 that the landing phase involves no more than two units. It is then necessary to suppress one of the balancing constraints in order to end the sampling. If we choose to relax the balancing constraint x^0 , the balancing equations (13) are maintained during the whole sampling process.

Since the balancing constraint x^0 is maintained during the whole sampling process, except perhaps for the last step, the imputation variance will be considerably reduced, though perhaps not totally eliminated. Also, the proposed

imputation method may be readily extended to the case of a categorical variable y with Q possible characteristics. The population U^* is then constituted of $n_m \times Q$ cells, each column being associated to one of the possible characteristics of y . The random balanced imputation process then follows the same lines as described above, each non-respondent i being given the j -th characteristic of the variable y if the cell (i, j) is selected in s^* .

5 The proposed algorithm

In this section, we describe and study a general algorithm for balanced random imputation, adapted from the Cube method originally proposed by Deville and Tillé (2004). This algorithm is related to that proposed by Chauvet (2009) in the context of stratified balanced sampling. Before introducing the steps of the algorithm, we introduce further notation. The population U^* is partitioned into n_m strata $U_1^*, \dots, U_i^*, \dots, U_{n_m}^*$ of equal size. The unit j in stratum U_i^* is associated to the cell (i, j) , that is, to the couple formed by the i -th non respondent and the j -th respondent. The inclusion probability and the value of the vector of balancing variables for unit j in stratum U_i^* are respectively given by ψ_{ij} and $\tilde{\mathbf{x}}_{ij}$ (see section 4). Let

$$A = \left(\frac{\tilde{\mathbf{x}}_{11}}{\psi_{11}}, \dots, \frac{\tilde{\mathbf{x}}_{ij}}{\psi_{ij}}, \dots, \frac{\tilde{\mathbf{x}}_{n_m n_r}}{\psi_{n_m n_r}} \right)$$

be a $(n_m + 1) \times (n_m n_r)$ matrix called the matrix of constraints. Also, let A_i be the $2 \times n_r$ submatrix obtained from A by taking the two lines i and $n_m + 1$, and the n_r columns $(i - 1) n_r + 1, \dots, i n_r$. That is, the submatrix A_i is obtained by taking the columns associated to units in U_i^* , and the lines associated to the constraints of (i) fixed-sample size in U_i^* and (ii) balancing on variable x^0 . Also, let $\psi_i = (\psi_{i1}, \dots, \psi_{in_r})'$ denote the vector of inclusion probabilities for units in the stratum U_i^* . To fix ideas, consider the particular case where $n_m = 3$ and $n_r = 4$. The matrix A may then be written as

$$A = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 & \dots & 1 \\ \frac{x_{11}^0}{\psi_{11}} & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \frac{x_{34}^0}{\psi_{34}} \end{pmatrix},$$

and the matrix A_1 is given by

$$A_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ \frac{x_{11}^0}{\psi_{11}} & \dots & \dots & \frac{x_{14}^0}{\psi_{14}} \end{pmatrix}.$$

The two first phases of the imputation process are given in Algorithm 1. During the first phase of Algorithm 1, independent flight phases are performed inside each stratum U_i^* , $i = 1, \dots, n_m$. The choice of the vector $u_i(t)$ in Sub-step 1 implies that the balancing equations for U_i^* remain exactly respected,

Phase 1 : for $i = 1, \dots, n_m$, initialize with $\psi_i(0) = \psi_i$. Then at any step $t = 1, \dots, T_i$, do:

Sub-step 1: Generate any $n_r \times 1$ vector $u_i(t) = (u_{i1}(t), \dots, u_{in_r}(t))'$ such that (i) $u_i(t)$ is in the kernel of A_i and (ii) $u_{ij}(t) = 0$ if $u_{ij}(t-1)$ is an integer.

Sub-step 2: Compute $\lambda_{1i}^*(t)$ and $\lambda_{2i}^*(t)$ the largest values of $\lambda_{1i}(t)$ and $\lambda_{2i}(t)$ such that $0 \leq \lambda_{1i}(t), \lambda_{2i}(t) \leq 1$. Note that $\lambda_{1i}^*(t) > 0$ and $\lambda_{2i}^*(t) > 0$.

Sub-step 3: Select

$$\psi_i(t) = \begin{cases} \psi_i(t-1) + \lambda_{1i}^*(t) u_i(t) & \text{with probability } q_i(t) \\ \psi_i(t-1) - \lambda_{2i}^*(t) u_i(t) & \text{with probability } 1 - q_i(t) \end{cases}$$

where $q_i(t) = \lambda_{2i}^*(t) / (\lambda_{1i}^*(t) + \lambda_{2i}^*(t))$.

Phase 2 : let

$$\phi(0) = (\psi'_1(T_1), \dots, \psi'_i(T_i), \dots, \psi'_{n_m}(T_{n_m}))'$$

and

$$v(0) = (u'_1(T_1), \dots, u'_i(T_i), \dots, u'_{n_m}(T_{n_m}))'$$

Then at any step $t = 1, \dots, T$, do:

Sub-step 1: Generate any vector $v(t) = (v_{11}(t), \dots, v_{ij}(t), \dots, v_{n_m n_r}(t))'$ such that (i) $v(t)$ is in the kernel of A and (ii) $v_{ij}(t) = 0$ if $v_{ij}(t-1)$ is an integer.

Sub-step 2: Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$ the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ such that $0 \leq \lambda_1(t), \lambda_2(t) \leq 1$. Note that $\lambda_1^*(t) > 0$ and $\lambda_2^*(t) > 0$.

Sub-step 3: Select

$$\phi(t) = \begin{cases} \phi(t-1) + \lambda_1^*(t) v(t) & \text{with probability } q(t) \\ \phi(t-1) - \lambda_2^*(t) v(t) & \text{with probability } 1 - q(t) \end{cases}$$

where $q(t) = \lambda_2^*(t) / (\lambda_1^*(t) + \lambda_2^*(t))$.

Algorithm 1 Cube method for balanced random imputation

that is, we obtain fixed-size sampling and balancing on variable x^0 in each stratum U_i^* . The choice of $\lambda_{1i}^*(t)$ and $\lambda_{2i}^*(t)$ in Sub-step 2 implies that the vector $\psi_i(t)$ has at least one more integer component than $\psi_i(t-1)$. This means that at each step t , one more unit is either sampled or definitely rejected. Finally, the random choice in Sub-step 3 implies that the inclusion probabilities are also exactly respected. This phase stops when it is no longer possible to select a vector $u_i(t)$ such that both (i) and (ii) are satisfied. T_i denotes the time when the flight phase stops in stratum U_i^* .

During the second phase of Algorithm 1, the condition of balancing on variable x^0 for any stratum U_i^* is replaced by a global condition of balancing on variable x^0 for the population U^* . The condition of fixed-size sampling inside each stratum U_i^* is maintained, since one residual exactly needs to be selected for any non respondent i . The second phase follows the same lines as in phase 1: at each step t , a random choice is performed such that both the balancing equations and the inclusion probabilities remain exactly respected while one more unit is either sampled or definitely rejected. This phase stops when it is no more possible to select a vector $v(t)$ such that both (i) and (ii) are satisfied. T denotes the time when the flight phase stops in phase 2.

At the end of phase 2, this is no more possible for the balancing condition on variable x^0 to hold exactly. If some units are neither selected nor rejected at the end of phase 2, a last, landing phase is involved, where a sampling design among the r remaining units is defined so that the inclusion probabilities are still exactly respected. Proposition 4 of Deville and Tillé (2004, p. 902) states that

$$\left| \sum_{(i,j) \in s^*} \frac{x_{ij}^0}{\psi_{ij}} - \sum_{(i,j) \in U^*} x_{ij}^0 \right| \leq r \times \max_{(i,j) \in U^*} \left| \frac{x_{ij}^0}{\psi_{ij}} \right|. \quad (16)$$

Equation (16) implies that the balancing condition (15) will be closely respected if the number r of units that remain after phases 1 and 2 is small. Proposition 1 in Deville and Tillé (2004, p. 897) states that r is no greater than the number of balancing variables, which equals $n_m + 1$. A smaller upper bound for r is given by Theorem 1.

Theorem 1 *The number r of units (i, j) such that $u_{ij}(T)$ is not an integer at the end of phase 2 of Algorithm 1 is no greater than 2.*

The proof is given in Appendix. The balancing condition (15) is thus preserved during the whole sampling process, except perhaps in the random choice for the last two units. The simulation results in Chauvet et al (2010) confirms that the imputation variance is almost completely eliminated if balanced random imputation is used.

6 Conclusion

In this paper, an adaptation of the Cube algorithm is proposed for the purpose of balanced random imputation. It can be applied to any type of regression method, under any type of sampling design and for both continuous and categorical variables. We demonstrated that the proposed algorithm enables the random selection of residuals while satisfying exactly, or very closely, a balancing constraint. Consequently, the variance imputation may be expected to be significantly reduced.

A Proof of Theorem 1

We first show that, at the end of phase 1 of Algorithm 1, there are only two possibilities for a given stratum U_i^* : either the vector $u_i(T_i)$ has two non-integer components, either the vector $u_i(T_i)$ has only integer components. It follows from Proposition 1 of Deville and Tillé (2004, p. 897) that the number r_i of non-integer components of $u_i(T_i)$ is lower than 2, that is, to the number of balancing variables used in the stratum U_i^* in phase 1. Since $u_i(T_i)$ is in the kernel of A_i , we have

$$\sum_{j=1}^{n_r} u_{ij}(T_i) = 0. \quad (17)$$

From equation (17), it follows that the vector $u_i(T_i)$ may not have only one non-integer component.

Now, let $r_i(0) = r_i$, and let $J(0)$ denote the number of strata U_i^* such that $r_i > 0$ at the end of phase 1. Similarly, let $r_i(T)$ denote the number of non-integer components of the sub-vector

$$v_i(T) = (v_{i1}(T), \dots, v_{in_r}(T))',$$

and let $J(T)$ denote the number of strata U_i^* such that $r_i(T) > 0$ at the end of phase 2. We have $r_i(T) \leq r_i(0) \leq 2$. Since $v(T)$ is in the kernel of A , $v_i(T)$ is in the kernel of A_i and we have

$$\sum_{j=1}^{n_r} v_{ij}(T) = 0. \quad (18)$$

From equation (18), it follows that the vector $v_i(T)$ may not have only one non-integer component. Consequently, at the end of phase 2, the vector $v(T)$ has exactly $2 \times J(T)$ non-integer components. On the other hand, there remains $J(T) + 1$ balancing constraints at the end of phase 2: one constraint of fixed size for any stratum U_i^* such that $r_i(T) > 0$, and the balancing constraint on variable x^0 . Since T denotes the time when the flight phase stops, it follows from Result 34 of Tillé (2006, p. 153) that

$$2 \times J(T) \leq J(T) + 1,$$

and consequently $J(T) \leq 1$. Since there remains at most one stratum U_i^* with $r_i(T) > 0$, and since $r_i(T) \leq 2$, the result follows.

References

- Chauvet G (2009) Stratified balanced sampling. *Survey Methodology* 35:115–119
- 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
- Chauvet G, Deville JC, Haziza D (2010) On balanced random imputation in surveys
- Deville JC, Tillé Y (2004) Efficient balanced sampling: the cube method. *Biometrika* 91(4):893–912
- Kim J, Fuller W (2004) Fractional hot-deck imputation. *Biometrika* 91:559–578
- Tillé Y (2006) Sampling algorithms. Springer Series in Statistics, Springer, New York