On Balanced Random Imputation in Surveys

Guillaume Chauvet, Jean-Claude Deville and David Haziza[1]

ABSTRACT

Random imputation methods are often used in practice because they tend to preserve the distribution of the variable being imputed, which is an important property when the goal is to estimate quantiles. Also, random hot-deck imputation, which is a random imputation method, is often used if the variable being imputed is categorical because it eliminates the possibility of impossible values. However, random imputation methods introduce an additional amount of variability, called the imputation variance, due to the random selection of residuals. In this paper, adapting the Cube method (Deville and Tillé, 2004) for selecting balanced samples we propose a class of random balanced imputation methods which reduce/eliminate the imputation variance while preserving the distribution of the variable being imputed. A limited simulation study supports our finding.

KEYWORDS: Balanced sampling; deterministic imputation; distribution function; imputation variance; random imputation;

[1] Guillaume Chauvet (chauvet@ensai.fr) and Jean-Claude Deville (deville@ensai.fr), Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France.

David Haziza (David.Haziza@umontreal.ca), Département de mathématiques et de statistique, Université de Montréal, Montréal, Québec, H3C 3J7, Canada.

1. INTRODUCTION

To compensate for item nonresponse in surveys, imputation methods are often used. The latter are used to replace missing values with artificial values in order to reduce, as much as possible, the bias and the variance introduced because of the missing values. Imputation methods may be classified into two broad classes: deterministic and random (or stochastic). Deterministic methods are those that yield a fixed imputed value given the sample if the imputation process is repeated as opposed to random methods that do not necessarily yield the same imputed value. One popular random imputation method used in practice is random hot-deck imputation that consists of selecting respondent (donor) values from the set of respondents to impute the missing values. In practice, it is often required to estimate population totals (or means) or quantiles (such as the median). While deterministic imputation methods lead to asymptotically unbiased estimators if the underlying imputation or nonresponse model is correctly specified (e.g., Haziza, 2008), they are not appropriate when the objective is to estimate a quantile (e.g., a median) because this type of imputation methods tends to distort the distribution of the variables being imputed. As a result, estimators of quantiles could be severely biased, especially if the nonresponse rate is appreciable. To preserve the distribution, it is customary to use a random imputation method. Also, if the variable being imputed is categorical, random hot-deck is preferable to avoid the possibility of impossible values in the imputed data file. However, this type of imputation methods introduces an additional amount of variability (called the imputation variance) due to the random selection of residuals. In some cases, the contribution of the imputation variance is important resulting in potentially inefficient estimators. It is thus desirable to develop imputation strategies which considerably reduces (or eliminates) the imputation variance while preserving the distribution of the variable being imputed.

In the literature, three general approaches have been considered for reducing the imputation variance. First, the fractional imputation approach, which consists of replacing each missing value with $M \geq 2$ imputed values selected randomly and assigning a weight to each imputed value. For example, each imputed value may receive 1/*M* times the original weight. Fractional imputation was originally proposed by Kalton and Kish (1981, 1984) and studied by Fay (1996), Kim and Fuller (2004) and Fuller and Kim (2005). It is similar to multiple imputation (Rubin, 1987), although the estimation procedures are different. It can be shown that the imputation variance decreases as *M* increases. One drawback of fractional imputation is that it may be cumbersome in practice since *M* imputed values are needed for each missing value. Also, the vast majority of surveys use single imputation methods. The second approach consists of first imputing the missing values using a standard random imputation method (e.g., random hot deck imputation) and adjusting the imputed values in such a way that the imputation variance is eliminated. This approach was considered by Chen, Rao and Sitter (2000) in the case of random hot-deck imputation. One drawback of the method is that, once the imputed file is produced with the use of random hot-deck imputation, the imputed values need to be adjusted by the data user, which may be seen as not practical. Also, the adjustment procedure will generally lead to impossible imputed values in the case of categorical variables. Finally, the third approach consists of randomly selecting donors (or residuals) in such a way that the imputation variance is reduced. This approach was originally considered by Kalton and Kish (1981; 1984) in the context of simple random sampling who suggested that donors (or residuals) may be selected by stratified sampling within imputation classes or by systematic sampling from a list of respondents ordered by their value taken by the variable being imputed. The idea behind these types of procedures is to select imputed values so that appropriate balancing equations are (approximately) satisfied. Following Kalton and Kish,

Deville (2006) proposed an algorithm for selecting imputed values while satisfying appropriate balancing constraints.

In this paper, we propose a class of random imputation method which we call *balanced random imputation*, and which is closely related to the third approach advocated by Kalton and Kish (1981, 1984) and Deville (2006). We introduce a general algorithm for balanced random imputation, adapted from the Cube method originally proposed by Deville and Tillé (2004). The proposed method consists of randomly selecting donors (or residuals) while satisfying given constraints. It can be readily 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. We show that the proposed class of imputation methods has the advantage of reducing the imputation variance significantly while preserving the distribution of the variable being imputed.

In our view, the third approach is attractive because it uses single imputation to compensate for the missing values, which leads to the creation of a single data file. Also, once the data file is produced, the usual estimation methods can be readily applied by users. In other words, no special adjustments need to be made. Finally, even though the primary objective is to estimate population totals, analysts may also be interested in studying the distribution of the variables that have been imputed, in which case deterministic methods would generally lead to misleading inferences. For this reason, we advocate for the use of balanced imputation methods.

The outline of the paper is as follows: in section 2, we present the imputation model and the corresponding imputed estimator. We derive the expression of the imputation variance and

introduce the concept of balanced random imputation methods. A general algorithm, adapted from the so-called Cube method proposed by Deville and Tillé (2004), is presented in section 3. In section 4, we show that the estimated distribution function based on observed and imputed values is a consistent estimator of the true distribution function under balanced random imputation. The estimation of more complex functions is considered in section 5. In section 6, the case of a categorical variable is considered. A limited simulation study comparing several imputation methods in terms of relative efficiency is presented in section 7. Finally, we conclude in section 8 and describe some future work.

2. BALANCED RANDOM IMPUTATION

Let $U = \{1, 2, ..., N\}$ 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, $i = 1, ..., N$. In section 2 and 3, we consider the case of continuous *y*. The case of binary *y* is considered in section 6. 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 basic estimator is the expansion estimator given by

$$\hat{Y}_\pi = \sum_{i \in s} w_i y_i. \tag{2.1}$$

The estimator $\hat{Y}_\pi$ in (2.1) is p-unbiased for Y; that is, $E_p\left(\hat{Y}_\pi\right) = Y$, where the subscript *p* indicates the sampling design $p(s)$. 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

$$\hat{Y}_I = \sum_{i \in s} w_i r_i y_i + \sum_{i \in s} w_i \left(1 - r_i\right) y_i^*, \tag{2.2}$$

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.

Most of the imputation methods used in practice can be motivated by the general model

$$y_i = f(\mathbf{z}_i) + \sigma \sqrt{v_i}\, \varepsilon_i, \tag{2.3}$$

where $f(.)$ is a given function, $\mathbf{z} = \left(z_1, ..., z_K\right)'$ is a K-vector of auxiliary variables available at the imputation stage for all the sampled units, σ^2 is an unknown parameter and v_i is a known constant. The ε_i's denote independent and identically distributed random variables from a common law (*L*) with mean 0 and variance 1 The subscript *m* in (2.3) indicates that the expectations, variances and covariances are evaluated with respect to the model. The model (2.3) is often called an imputation model (e.g., Särndal, 1992).

In the case of deterministic imputation, the imputed value y_i^* is obtained by estimating $f(.)$ by $\hat{f}_r(.)$ using the responding units; that is, $y_i^* = \hat{f}_r\left(\mathbf{z}_i\right)$, for $i \in s_m$. Random imputation can be seen as a deterministic imputation to which a random noise ε_i^* is added. That is,

$$y_i^* = \hat{f}_r\left(\mathbf{z}_i\right) + \hat{\sigma} \sqrt{v_i}\, \varepsilon_i^*, \quad \text{for } i \in s_m, \tag{2.4}$$

where $\hat{\sigma}$ is an estimator of σ. In other words, for each random imputation, there is a corresponding deterministic imputation, which is obtained by setting $\varepsilon_i^* = 0$ for all *i*. The random quantity ε_i^* can be generated from a given distribution. However, in practice, it is

natural to select (usually with replacement) the random component ε_i^* from the set, $E_r = \{\tilde{e}_j ; j \in s_r\}$, of standardized residuals observed from the responding units, with probabilities

$$P(\varepsilon_i^* = \tilde{e}_j) = \omega_j / \sum_{l \in s} \omega_l r_l, \tag{2.5}$$

where $\tilde{e}_j = e_j - \bar{e}_r$, $e_j = \frac{1}{\hat{\sigma}\sqrt{v_j}}\left(y_j - \hat{f}_r(\mathbf{z}_i)\right)$, $\bar{e}_r = \sum_{j \in s} \omega_j r_j e_j / \sum_{j \in s} \omega_j r_j$ and ω_j is an imputation weight attached to unit *j*. This method for selecting the random residuals ε_i^* is nonparametric in nature since it consists of generating random residuals from the empirical distribution function of the residuals,

$$\hat{F}_{\varepsilon,r}(t) = \sum_{j \in s} \tilde{\omega}_j r_j \mathbf{1}(\tilde{e}_j \le t), \tag{2.6}$$

based on the responding units, where $\tilde{\omega}_j = \omega_j / \sum_{l \in s} \omega_l r_l$ and $\mathbf{1}(.)$ is the usual indicator function.

Several choices of ω_j are possible: 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. Other choices of imputation weights are possible (Haziza, 2009). Note that the imputed value y_i^* in (2.4) can be viewed as the sum of a deterministic component, $\hat{f}_r(\mathbf{z}_i)$, and a random component ε_i^*.

Letting $f(\mathbf{z}_i) = \mathbf{z}_i'\boldsymbol{\beta}$ in (2.3) leads to the model underlying (deterministic and random) regression imputation, where $\boldsymbol{\beta}$ is a K-vector of unknown parameters. Random regression imputation is thus obtained from (2.4) by setting $\hat{f}_r(\mathbf{z}_i) = \mathbf{z}_i'\hat{\mathbf{B}}_r$, where

$$\hat{\mathbf{B}}_r = \left[\sum_{i\in s} \omega_i r_i \mathbf{z}_i \mathbf{z}_i' / \nu_i\right]^{-1} \sum_{i\in s} \omega_i r_i \mathbf{z}_i y_i / \nu_i$$

is the weighted least square estimator of $\boldsymbol{\beta}$ based on the responding units. That is, random regression imputation uses the imputed values

$$y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r + \hat{\sigma}\sqrt{\nu_i}\varepsilon_i^*. \tag{2.7}$$

Random hot-deck imputation within classes, which is a popular method in practice, can be viewed as a special case of (2.7). It consists of first partitioning the sample into *K* imputation classes, $s_1, ..., s_k, ..., s_K$. 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; $k = 1, 2, ..., K.$ Random hot deck imputation (RHDI) within classes is obtained from (2.7) by setting $\mathbf{z}_i = \left(z_{1i}, ..., z_{Ki}\right)'$ and $\nu_i = \nu_k$ if i belong to class *k*.

Using the imputed values (2.4) in (2.2) leads to

$$\begin{aligned} \hat{Y}_I &= \sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i \left(1-r_i\right) \hat{f}_r\left(\mathbf{z}_i\right) + \sum_{i\in s} w_i \left(1-r_i\right) \hat{\sigma}\sqrt{\nu_i}\varepsilon_i^* \\ &= \sum_{i\in s} w_i r_i y_i + \sum_{i\in s} w_i \left(1-r_i\right) \hat{f}_r\left(\mathbf{z}_i\right) + \sum_{i\in s} w_i \left(1-r_i\right) \hat{\sigma}\sqrt{\nu_i} \sum_{j\in s} r_j d_{ji} \tilde{e}_j, \end{aligned} \tag{2.8}$$

where $d_{ji} = \begin{cases} 1 & \text{if the residual } \tilde{e}_j \text{ was selected for imputing the missing value } y_i \\ 0 & \text{otherwise} \end{cases}$

Let the subscript *q* indicate the nonresponse mechanism and the subscript *I* indicate the imputation mechanism. The total variance of $\hat{Y}_I$ given in (2.3) can be expressed as

$$V\left(\hat{Y}_I\right) = V_p E_q E_I\left(\hat{Y}_I \mid s\right) + E_p V_q E_I\left(\hat{Y}_I \mid s\right) + E_p E_q V_I\left(\hat{Y}_I \mid s\right). \tag{2.9}$$

The first term on the right hand side of (2.9) is the sampling variance, the second term is the nonresponse variance, whereas the third term is the imputation variance. The imputation variance is given by

$$E_p E_q V_I\left(\hat{Y}_I \mid s\right) = E_p E_q \left[\frac{\sum_{i \in s} w_i^2 \left(1 - r_i\right) \nu_i}{\sum_{i \in s} \omega_i r_i} \sum_{i \in s} \omega_i r_i \tilde{e}_i^2 \right]. \tag{2.10}$$

Under mild regularity conditions, the imputation variance given by (2.10) is $O\left(N^2/n\right)$, which is the same order of magnitude as the sampling and nonresponse variances. From (2.10), we note that the magnitude of the imputation variance will be small if (i) the response rate is high in which case the term $\frac{\sum_{i \in s} w_i^2 \left(1 - r_i\right) \nu_i}{\sum_{i \in s} \omega_i r_i}$ is likely to be small and (ii) the imputation model fits the data well, in which case the term $\sum_{i \in s} \omega_i r_i \tilde{e}_i^2$ will be small. Otherwise, the contribution of the imputation variance to the total variance can be appreciable.

For example, consider the case of simple linear regression imputation (SLRI). Deterministic SLRI is obtained from (2.7) by setting $\mathbf{z}_i = \left(1, z_i\right)'$, $\nu_i = 1$ and $\varepsilon_i^* = 0$ for all *i*. Random SLRI is obtained from (2.7) by setting $\mathbf{z}_i = \left(1, z_i\right)'$ and $\nu_i = 1$. Let $V_D\left(\hat{Y}_I\right)$ and $V_R\left(\hat{Y}_I\right)$ denote the total variance of $\hat{Y}_I$ under deterministic and random SLRI, 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 relative contribution of the imputation variance to the total variance, $C = \frac{V_R\left(\hat{Y}_I\right) - V_D\left(\hat{Y}_I\right)}{V_D\left(\hat{Y}_I\right)}$, can be approximated by

$$C \approx \frac{p(1-p)\left(1-\rho_{yz}^2\right)}{1-(1-p)\rho_{yz}^2}, \tag{2.11}$$

provided the sample size n is sufficiently large, where ρ_{yz} denotes the coefficient of correlation between y and z. Figure 1 shows the contribution (in %) of the imputation variance to the total variance for a fixed value of ρ_{yz} ($\rho_{yz} = 0.8$), whereas Figure 2 shows its contribution (in %) for a fixed value of p ($p = 0.5$). 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 (2.11) is maximum, and decreases in the interval $(p_{\max}, 1]$. Note that when $\rho_{yz} = 0.8,$ we have $p_{\max} \approx 0.51.$ Also, it is clear from Figure 2 that the contribution of the imputation variance decreases as the coefficient of correlation between y and z increases, as expected.

Figure 1: Contribution (in %) of the imputation variance to the total variance with $\rho_{yz} = 0.8$.

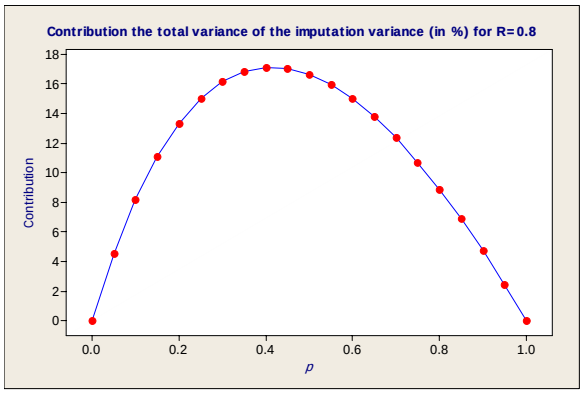
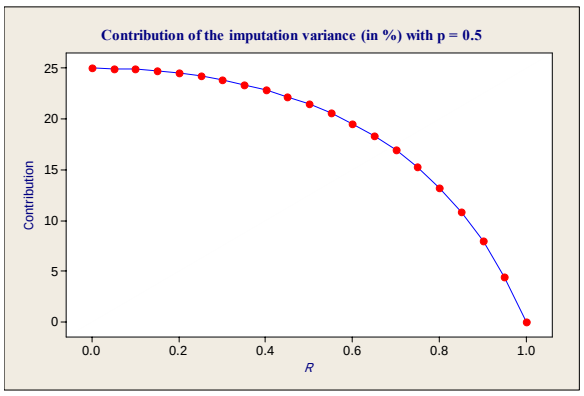


Figure 2: Contribution (in %) of the imputation variance to the total variance with $p = 0.5$.



We propose a balanced random imputation method which consists of selecting the residuals ε_i^* so that the following equation is (approximately) satisfied:

$$\sum_{i \in s} w_i (1 - r_i) \sqrt{v_i} \sum_{j \in s} r_j d_{ji} \tilde{e}_j = 0 . \qquad (2.12)$$

If the equation (2.12) is exactly satisfied, then the imputation variance is completely eliminated and the resulting estimator is fully efficient (Kim and Fuller, 2004). In some situations, it is not possible to satisfy (2.12) exactly but only approximately. In this case, the imputation variance is not completely eliminated but it is expected to be significantly reduced. In section 3, we describe a general algorithm for implementing the proposed imputation method.

In the special case of random hot-deck imputation within classes, the condition (2.12) reduces to

$$\bar{y}_{mk}^* = \bar{y}_{rk}, \quad k = 1, ..., K \qquad (2.13)$$

noting that $\sum_{j\in s_k} d_{ji} = 1$, where $\bar{y}_{rk} = \dfrac{\sum_{i\in s_k} \omega_i r_i y_i}{\sum_{i\in s_k} \omega_i r_i}$ is the weighted mean of the respondents in class *k* and $\bar{y}^*_{mk} = \dfrac{\sum_{i\in s_k} \omega_i (1-r_i) y^*_i}{\sum_{i\in s_k} \omega_i (1-r_i)}$ is the weighted mean of the imputed values in class *k*. In other words, eliminating the imputation variance will consist in selecting the imputed values at random within each class so that their mean matches the mean of the respondents within the same class. Chen, Rao and Sitter (2000) proposed a method for eliminating the imputation variance which consists of adjusting the imputed values obtained under RDHI so that (2.13) is satisfied. Note that our proposed balanced random imputation method does not require an adjustment of the imputed values. Rather, we select the imputed values at random so that (2.13) is satisfied, which is more attractive from a data user's perspective.

3. THE ALGORITHM

In this section, we propose a general algorithm for balanced random imputation, adapted from the Cube method originally proposed by Deville and Tillé (2004). Consider the $n_m \times n_r$ table below:

	1	...	j	...	n_r
1	$(\psi_{11}, \tilde{e}_1)$	...	$(\psi_{1j}, \tilde{e}_j)$	...	$(\psi_{1n_r}, \tilde{e}_{n_r})$
⋮	⋮		⋮		⋮
i	$(\psi_{i1}, \tilde{e}_1)$		$(\psi_{ij}, \tilde{e}_j)$		$(\psi_{in_r}, \tilde{e}_{n_r})$
⋮	⋮		⋮		⋮
n_m	$(\psi_{n_m1}, \tilde{e}_1)$	...	$(\psi_{n_mj}, \tilde{e}_j)$	...	$(\psi_{n_mn_r}, \tilde{e}_{n_r})$

where each cell (i, j) is given the value of the centered residual $\tilde{e}_j$ and the probability of selection $\psi_{ij} = \omega_j \Big/ \sum_{l\in s} \omega_l r_l$. Let U^* denote the population of $n_m \times n_r$ cells. Note that a random

imputation obtained from (2.4) may alternatively be seen as the without replacement selection of a random sample s^* of n_m cells in U^*. Since one residual exactly has to be selected for each nonrespondent, exactly one cell per row should be selected in s^*. Also, since the selection probabilities given by (2.5) have to be exactly satisfied, the cell (i, j) should be included in the sample with probability ψ_{ij}. If these two constraints are satisfied, then the random selection of the sample of cells s^* leads to the imputed values given by (2.4).

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

$$\sum_{j=1}^{n_r} d_{i^* j} = 1, \qquad i^* = 1, \cdots, n_m, \tag{3.1}$$

with $\sum_{j=1}^{n_r} d_{i^* j} = \sum_{(i,j)\in s^*} \delta_{i^* i}$, where $\delta_{i^* i} = 1$ if $i^* = i$ and $\delta_{i^* i} = 0$, otherwise. Since the sum of the inclusion probabilities on row i^* is equal to 1, we have

$$\sum_{j=1}^{n_r} \psi_{i^* j} = 1 = \sum_{(i,j)\in U^*} \psi_{ij}\delta_{i^* i}, \ i^* = 1, \cdots, n_m. \tag{3.2}$$

It follows that the system (3.1) can be written as a system of n_m balancing equations

$$\sum_{(i,j)\in s^*} \frac{\mathbf{x}_{ij}}{\psi_{ij}} = \sum_{(i,j)\in U^*} \mathbf{x}_{ij} \tag{3.3}$$

on a n_m vector of variables

$$\mathbf{x} = \left(x^1, \cdots, x^{n_m}\right)', \tag{3.4}$$

where the variable x^{i^*} takes the value $x_{ij}^{i^*} = \psi_{ij}\delta_{i^* i}$ on the cell (i, j). The selection of a sample of cells with respect to the balancing equations (3.3) and prescribed inclusion probabilities ψ_{ij} may be handled with balanced sampling by means of the Cube method proposed by Deville and Tillé (2004). The Cube algorithm is described in the appendix. Note that the selection of

s^* with the algorithm proposed and balancing variables $\mathbf{x}$ given in (3.4) is equivalent to selecting independently and with replacement the random components ε_i^* in (2.4) from the set $E_r = \left\{\tilde{e}_j ; j \in s_r\right\}$. In other words, the traditional imputation method in (2.4) can be viewed as a special case of the proposed method.

Now, observe that (2.12) can also be written as a system of balancing equations. The left hand side (2.12) may be written as

$$\sum_{i \in s_m} w_i \sqrt{v_i} \sum_{j \in s_r} d_{ji} \tilde{e}_j = \sum_{(i,j) \in s^*} \frac{x_{ij}^0}{\psi_{ij}}, \tag{3.5}$$

with $x_{ij}^0 = w_i \sqrt{v_i} \psi_{ij} \tilde{e}_j$ for the cell (i, j). On the other hand, we have

$$\begin{aligned} \sum_{(i,j) \in U^*} x_{ij}^0 &= \sum_{i \in s_m} w_i \sqrt{v_i} \sum_{j \in s_r} \psi_{ij} \tilde{e}_j \\ &= \sum_{i \in s_m} w_i \sqrt{v_i} \frac{\sum_{j \in s_r} \omega_j \tilde{e}_j}{\sum_{j \in s_r} \omega_j} \\ &= 0 \end{aligned}$$

since $\sum_{j \in s_r} \omega_j \tilde{e}_j = 0.$ It follows that (2.12) may be written as

$$\sum_{(i,j) \in s^*} \frac{x_{ij}^0}{\psi_{ij}} = \sum_{(i,j) \in U^*} x_{ij}^0 . \tag{3.6}$$

Selecting a sample balanced on variables $\tilde{\mathbf{x}} = \left(x^0, \mathbf{x}'\right)'$, where $\mathbf{x}$ is given in (3.4), with inclusion probabilities ψ_{ij}, ensures that (i) the selection probabilities given in (2.5) are exactly satisfied, (ii) one residual $\tilde{e}_j$ exactly is selected for each missing value y_i, and (iii) equation (2.12) is exactly satisfied and, as a result, the variance imputation is eliminated.

In practice, note that there may exist no sample s^* such that both equations (3.3) and (3.6) are exactly satisfied. The Cube method then involves a rounding process called the landing phase (see the Appendix) in order to end the sampling. If such a situation occurs, the inclusion probabilities remain exactly respected and exactly one cell is selected in each row, but equation (3.6) will only be approximately satisfied, in which case, the imputation variance will be considerably reduced but not totally eliminated.

4. CONSISTENCY OF THE DISTRIBUTION FUNCTION

In this section, we show that the proposed balanced random imputation preserves the distribution of the variable being imputed. The finite population distribution function can be written as $F_N(t) = \frac{1}{N}\sum_{i\in U}\mathbf{1}(y_i \le t)$. A complete data estimator of $F_N(t)$ is given by

$$\hat{F}_N(t) = \sum_{i\in s}\tilde{w}_i\mathbf{1}(y_i \le t), \tag{4.1}$$

where $\tilde{w}_i = w_i \Big/ \sum_{l\in s} w_l$. An imputed estimator of $F_N(t)$ is given by

$$\hat{F}_I(t) = \sum_{i\in s}\tilde{w}_i r_i \mathbf{1}(y_i \le t) + \sum_{i\in s}\tilde{w}_i (1-r_i)\mathbf{1}(y_i^* \le t). \tag{4.2}$$

We consider the case of weighted random imputation; that is, $\omega_i \equiv w_i$. We assume that the following regularity conditions hold:

C1: max $w_i = O(N/n)$;

C2: There exists a constant κ such that $\kappa < p_i \equiv P(r_i = 1)$ for all i;

C3: $\sup_t\left(\left|\hat{F}_{\varepsilon,r}(t) - F_\varepsilon(t)\right|\right) = O_p\left(n^{-1/2}\right)$;

C4 : $V_I\left(\sum_{(i,j)\in s^*}\frac{b_{ij}}{\psi_{ij}}\middle| s,s_r,\widehat{\mathbf{x}}\right)=\left(1+o_p(1)\right)V_I^{app}\left(\sum_{(i,j)\in s^*}\frac{b_{ij}}{\psi_{ij}}\middle| s,s_r,\widehat{\mathbf{x}}\right)$, where $V_I\left(.\middle| s,s_r,\widehat{\mathbf{x}}\right)$ denotes the variance, conditional on s and s_r, under imputation by means of balanced sampling with inclusion probabilities ψ_{ij} and balancing variables $\widehat{\mathbf{x}}$, b_{ij} denotes the value taken by a non-random (conditional on s and s_r) variable b in cell (i,j),

$$V_I^{app}\left(\sum_{(i,j)\in s^*}\frac{b_{ij}}{\psi_{ij}}\middle| s,s_r,\widehat{\mathbf{x}}\right)=\left(\sum_{(i,j)\in U^*}\psi_{ij}(1-\psi_{ij})\left(\frac{b_{ij}}{\psi_{ij}}-\frac{\hat{b}_{ij}(\widehat{\mathbf{x}})}{\psi_{ij}}\right)^2\right) \text{ and}$$

$$\hat{b}_{ij}(\widehat{\mathbf{x}})=\widehat{\mathbf{x}}_{ij}'\left(\sum_{(i',j')\in U^*}\psi_{i'j'}(1-\psi_{i'j'})\left(\frac{\widehat{\mathbf{x}}_{i'j'}}{\psi_{i'j'}}\right)\left(\frac{\widehat{\mathbf{x}}_{i'j'}}{\psi_{i'j'}}\right)'\right)^{-1}\left(\sum_{(i',j')\in U^*}\psi_{i'j'}(1-\psi_{i'j'})\left(\frac{\widehat{\mathbf{x}}_{i'j'}}{\psi_{i'j'}}\right)\left(\frac{b_{i'j'}}{\psi_{i'j'}}\right)'\right) \qquad (4.3)$$

denotes a weighted prediction of b_{ij}.

The assumption (C1) guarantees that no extreme weight dominates the others. The assumption C2 states that the response probability is bounded away from 0. The assumption C3 states that the empirical distribution of the residuals corresponding to the responding units is a consistent estimator of the true distribution of the errors. Finally, the assumption (C4) gives a variance approximation for balanced sampling analog to that considered in Deville and Tillé (2005). Note that assumption C4 was proved by Hajek (1964) in the special case of the maximum entropy balanced sampling design with $\widehat{\mathbf{x}}_{ij}=\widehat{x}_{ij}=\psi_{ij}$.

Theorem: Suppose that conditions (C1-C4) hold. If the imputation model (2.3) holds, then $\hat{F}_I(t)-F_N(t)=O_p\left(n^{-1/2}\right)$.

Proof: The total error of $\hat{F}_I(t)$, $\hat{F}_I(t)-F_N(t)$, can be expressed as

$$\hat{F}_I(t)-F_N(t)=\left[\hat{F}_N(t)-F_N(t)\right]+\left[\hat{F}_I(t)-\hat{F}_N(t)\right].$$

First, it follows from standard regularity conditions (e.g., Isaki and Fuller, 1982) that $\hat{F}_N(t) - F_N(t) = O_p\left(n^{-1/2}\right)$. It remains to show that $\hat{F}_I(t) - \hat{F}_N(t) = O_p\left(n^{-1/2}\right)$. To that end, let

$$\begin{aligned} \Delta \equiv \hat{F}_I(t) - \hat{F}_N(t) &= \sum_{i \in s} \tilde{w}_i (1 - r_i)\left[\mathbf{1}\left(\varepsilon_i^* \le \hat{t}_i\right) - \mathbf{1}\left(\varepsilon_i \le t_i\right)\right], \\ &= T_1 - T_2 \end{aligned} \tag{4.4}$$

where $T_1 = \sum_{i \in s} \tilde{w}_i (1 - r_i)\mathbf{1}\left(\varepsilon_i^* \le \hat{t}_i\right)$, $T_2 = \sum_{i \in s} \tilde{w}_i (1 - r_i)\mathbf{1}\left(\varepsilon_i \le t_i\right)$, $\hat{t}_i = \dfrac{t - \hat{f}_r(z_i)}{\hat{\sigma}\sqrt{v_i}}$ and $t_i = \dfrac{t - f(z_i)}{\sigma\sqrt{v_i}}$. We first show that the conditional nonresponse/imputation expectation of Δ in (4.4), given by

$$E(\Delta) = E_m E_I\left(\Delta \middle| s, s_r, \tilde{\mathbf{x}}\right) \tag{4.5}$$

is $O(n^{-1/2})$, where $\tilde{\mathbf{x}}$ denotes the vector of balancing variables used for imputation; see section 3. First note that T_2 is independent of the imputation mechanism. Consequently, we have

$$\begin{aligned} E_I\left(\Delta \middle| s, s_r, \tilde{\mathbf{x}}\right) &= E_I\left(T_1 - T_2 \middle| s, s_r, \tilde{\mathbf{x}}\right) \\ &= E_I\left(T_1 \middle| s, s_r, \tilde{\mathbf{x}}\right) - T_2 \\ &= \sum_{i \in s} \tilde{w}_i (1 - r_i) \hat{F}_{r,\varepsilon}\left(\hat{t}_i\right) - T_2 \\ &= U_1 + U_2 + U_3 \end{aligned} \tag{4.6}$$

where $U_1 = \sum_{i \in s} \tilde{w}_i (1 - r_i)\left[\hat{F}_{r,\varepsilon}\left(\hat{t}_i\right) - \hat{F}_{\varepsilon}\left(\hat{t}_i\right)\right]$, $U_2 = \sum_{i \in s} \tilde{w}_i (1 - r_i)\left[\hat{F}_{\varepsilon}\left(\hat{t}_i\right) - F_{\varepsilon}\left(t_i\right)\right]$, and $U_3 = \sum_{i \in s} \tilde{w}_i (1 - r_i)\left[F_{\varepsilon}\left(t_i\right) - \mathbf{1}\left(\varepsilon_i \le t_i\right)\right]$. As $\sum_{i \in s} \tilde{w}_i (1 - r_i)$ remains bounded under C1-C2, the term $E_m\left(U_1 \mid s, s_r\right)$ is $O\left(n^{-1/2}\right)$ under C3. We also have $E_m(U_3 \mid s, s_r) = 0$, and $V_m(U_3 \mid s, s_r) = \sum_{i \in s} \tilde{w}_i^2 (1 - r_i) F_{\varepsilon}\left(t_i\right)\left(1 - F_{\varepsilon}\left(t_i\right)\right)$ is asymptotically $O\left(n^{-1}\right)$ under C1-C3, so that $U_3 = O_p(n^{-1/2})$. Now, suppose that $f(z_i) = f\left(z_i, \boldsymbol{\beta}\right)$ and $\hat{f}_r(z_i) = f\left(z_i, \hat{\mathbf{B}}_r\right)$ such that

$\hat{\mathbf{B}}_r - \boldsymbol{\beta} = O_p\left(n^{-1/2}\right)$. Moreover, we assume that $F_\varepsilon(.)$ has bounded density. It follows that $E_m\left(U_2 \middle| s, s_r\right)$ is $O\left(n^{-1/2}\right)$. Consequently, $E_m E_I\left(\Delta \middle| s, s_r, \tilde{\mathbf{x}}\right)$ is $O\left(n^{-1/2}\right)$.

We now show that the conditional nonresponse/imputation variance of Δ in (4.4), given by

$$V(\Delta) = E_m V_I\left(\Delta \middle| s, s_r, \tilde{\mathbf{x}}\right) + V_m E_I\left(\Delta \middle| s, s_r, \tilde{\mathbf{x}}\right) \tag{4.7}$$

is $O\left(n^{-1}\right)$, where $\tilde{\mathbf{x}}$ denotes the vector of balancing variables used for imputation; see section 3. First, note that

$$\begin{aligned} V_I\left(\Delta \middle| s, s_r, \tilde{\mathbf{x}}\right) &= V_I\left(T_1 - T_2 \middle| s, s_r, \tilde{\mathbf{x}}\right) \\ &= V_I\left(T_1 \middle| s, s_r, \tilde{\mathbf{x}}\right) \end{aligned}, \tag{4.8}$$

since T_2 is independent of the imputation mechanism. Now, observe that T_1 may alternatively be written as

$$\begin{aligned} T_1 &= \sum_{i \in s} \tilde{w}_i (1 - r_i) \sum_{j \in s} r_j d_{ji} \mathbf{1}\left(\tilde{e}_j \le \hat{t}_i\right) \\ &= \sum_{(i,j) \in s^*} \frac{b_{ij}}{\psi_{ij}} \end{aligned}, \tag{4.9}$$

where $b_{ij} = \tilde{w}_i \psi_{ij} \mathbf{1}\left(\tilde{e}_j \le \hat{t}_i\right)$. Condition C4 implies that $V_I\left(T_1 \middle| s, s_r, \tilde{\mathbf{x}}\right)$ is asymptotically equivalent to $V_I^{app}\left(T_1 \middle| s, s_r, \tilde{\mathbf{x}}\right)$, and

$V_I^{app}\left(T_1 \middle| s, s_r, \tilde{\mathbf{x}}\right) \le V_I^{app}\left(T_1 \middle| s, s_r, \mathbf{x}\right)$ since $\tilde{\mathbf{x}}$ includes $\mathbf{x}$; see equation (3.4). The term $V_I^{app}\left(T_1 \middle| s, s_r, \mathbf{x}\right)$ is in turn asymptotically equivalent to $V_I\left(T_1 \middle| s, s_r, \mathbf{x}\right)$ by assumption C4, which is the imputation variance we obtain when the random quantities ε_i^* are selected independently and with replacement. Therefore, we get $V_I\left(T_1 \middle| s, s_r, \mathbf{x}\right) = \sum_{i \in s} \tilde{w}_i^2 (1 - r_i) \hat{F}_{r,\varepsilon}\left(\hat{t}_i\right)\left(1 - \hat{F}_{r,\varepsilon}\left(\hat{t}_i\right)\right)$. Under C1-C3, the order of magnitude of the latter term is $O\left(n^{-1}\right)$.

We now consider the second term on the right-hand size of (4.7). It follows from (4.6) that $E_I\left(\Delta\middle|s,s_r,\tilde{\mathbf{x}}\right)=T_3-T_2$, where $T_3=\sum_{i\in s}\tilde{w}_i(1-r_i)\hat{F}_{r,\varepsilon}(\hat{t}_i)$. Note that T_3 depends only on the ε's in the set of respondents s_r, whereas T_2 depends only on the ε's in the set of respondents s_m. As a result, these two terms are independent with respect to the imputation model (2.3), and $V_m E_I\left(\Delta\middle|s,s_r,\tilde{\mathbf{x}}\right)=V_m\left(T_3\middle|s,s_r\right)+V_m\left(T_2\middle|s,s_r\right)$. We have $V_m\left(T_2\middle|s,s_r\right)=\sum_{i\in s}\tilde{w}_i^2(1-r_i)F_\varepsilon(t_i)(1-F_\varepsilon(t_i))$. Under C1 and C2, the latter quantity is $O\left(n^{-1}\right)$.

Consequently, Δ is asymptotically $O_p(n^{-1/2})$ and the two quantities $\hat{F}_I(t)$ and $\hat{F}_N(t)$ can be taken as estimators of the same quantity, if the imputation model is correctly specified.

5. THE CASE OF A BINARY VARIABLE

In this section, we consider the case of a binary variable *y*. For simplicity, we first consider the case of a binary variable. Let $y_i=1$ if unit *i* possesses a given characteristic of interest and $y_i=0,$ otherwise. We assume that the *y*-variable is parametrically modeled; that is,

$$\phi_i\equiv P\left(y_i=1\right)=p\left(\mathbf{z}_i;\boldsymbol{\gamma}^0\right),$$

for some function $p\left(\mathbf{z}_i;.\right)$ with parameter $\boldsymbol{\gamma}$ evaluated at $\boldsymbol{\gamma}^0$, where $\mathbf{z}_i$ is a vector of auxiliary variables (as in section 2) available for both respondents and nonrespondents. Let $\hat{\boldsymbol{\gamma}}$ be an estimator of γ^0 and

$$\hat{\phi}_i=p\left(\mathbf{z}_i;\hat{\boldsymbol{\gamma}}\right) \tag{5.1}$$

be the estimated probability for unit i of possessing the characteristic of interest. We are interested in estimating the proportion of individuals that possess the characteristic, $\bar{Y} = \frac{1}{N}\sum_{i \in U} y_i$. A complete data estimator of $\bar{Y}$ is given by $\bar{y} = \hat{Y}_\pi / N$, where $\hat{Y}_\pi$ is given by (2.1), whereas an imputed estimator of $\bar{Y}$ is given by $\bar{y}_I = \hat{Y}_I / N$, where $\hat{Y}_I$ is given by (2.2). To impute binary variables, deterministic imputation methods (such as regression imputation) are usually rejected because they generally lead to impossible values in the data file. One notable exception is nearest-neighbour imputation that uses donor (observed) values to impute for missing values. In this section, we consider a version of RHDI (see section 2). For missing y_i, RHDI uses

$$y_i^* = \begin{cases} 1 & \text{with probability } \hat{\phi}_i \\ 0 & \text{with probability 1-}\hat{\phi}_i \end{cases}$$

In this context, the imputation variance of $\hat{Y}_I$ is given by

$$E_p E_q V_I\left(\hat{Y}_I \mid s\right) = E_p E_q \left[\frac{1}{N^2}\sum_{i \in s} w_i^2 \left(1-r_i\right)\hat{\phi}_i\left(1-\hat{\phi}_i\right)\right]. \qquad (5.2)$$

Under mild regularity conditions, the imputation variance in (5.2) is $O\left(n^{-1}\right)$, which is the same order of magnitude than the sampling and the nonresponse variances.

We propose a balanced random imputation method which consists of selecting the imputed values y_i^* so that the following equation is approximately satisfied

$$\sum_{i \in s} w_i (1-r_i) y_i^* = \sum_{i \in s} w_i (1-r_i)\hat{\varphi}_i \qquad (5.3)$$

If the equation (5.3) holds exactly, then the imputation variance is completely eliminated. We describe below an adaptation of the algorithm introduced in section 3 to handle the case of a binary variable. Consider the $n_m \times 2$ table

	1	2
1	(ψ_{11}, e_{11})	(ψ_{12}, e_{12})
$\vdots$	$\vdots$	$\vdots$
i	(ψ_{i1}, e_{i1})	(ψ_{i2}, e_{i2})
$\vdots$	$\vdots$	$\vdots$
n_m	$(\psi_{n_m 1}, e_{n_m 1})$	$(\psi_{n_m 2}, e_{n_m 2})$

where each cell (i, j) is given the value $e_{ij} = 1$ if $j = 1$ and $e_{ij} = 0$ if $j = 2$, and the probability $\psi_{ij} = \hat{\varphi}_i$ if $j = 1$ and $\psi_{ij} = 1 - \hat{\varphi}_i$ if $j = 2$. A random imputation for variable y may be seen as the random selection of a sample s^* in the population U^* of cells, in the sense that if the cell (i, j) is selected in s^*, the value 1 will be used for imputation of missing y_i if $j = 1$, and the value 0 will be used if $j = 2$. The selection of s^* must be such that (i) one cell exactly is selected in each line, (ii) the estimated probabilities in (5.1) are exactly respected and (iii) equation (5.3) holds approximately for the imputation variance to be strongly reduced. In a similar way than the method presented in section 3, the selection of such a sample s^* may be handled with the Cube method presented in the Appendix, with inclusion probabilities ψ_{ij} and the vector $\tilde{\mathbf{x}}_{ij} = \left(x_{ij}^0, x_{ij}^1, \cdots, x_{ij}^{n_m}\right)$, where $x_{ij}^0 = w_i \psi_{ij} e_{ij}$ and $x_{ij}^{i^*} = \psi_{ij} \delta_{i^* i}$ for $i^* = 1, \cdots, n_m$, where $\delta_{i^* i} = 1$ if $i^* = i$ and $\delta_{i^* i} = 0$, otherwise.

We now briefly consider the case of a categorical variable y with K possible characteristics. Let $y_i = 1$ if unit i possesses the first characteristic of interest, $y_i = 2$ if unit i possesses the

second characteristic of interest, and so on. Assume that the y variable is parametrically modeled; that is,

$$\varphi_i^k \equiv P\left(y_i = k\right) = p\left(\mathbf{z}_i; \gamma^{0,k}\right)$$

where $\sum_{k=1}^{K} \varphi_i^k = 1$, for some function $p\left(\mathbf{z}_i;.\right)$ with parameter $\boldsymbol{\gamma}$ evaluated at $\gamma^{0,k}$ where $\mathbf{z}_i$ is a vector of auxiliary variables available for both respondents and nonrespondents. Let $\hat{\gamma}^k$ be an estimator of $\gamma^{0,k}$ and

$$\hat{\varphi}_i^k = p\left(\mathbf{z}_i; \hat{\gamma}^k\right) \tag{5.4}$$

be the estimated probability for unit i of possessing the characteristic of interest k, where $\sum_{k=1}^{K} \hat{\varphi}_i^k = 1$. We are interested in estimating the proportion of individuals that possess the characteristic k, $\bar{Y}_k = \frac{1}{N}\sum_{i\in U} 1(y_i = k)$. An imputed estimator of $\bar{Y}_k$ is given by $\bar{y}_{kI} = \hat{Y}_{kI} / N$, where $\hat{Y}_{kI}$ is given by (2.2) where the variable y_i is replaced by $1(y_i = k)$. For missing y_i, we use

$$y_i^* = k \quad \text{with probability} \quad \hat{\varphi}_i^k.$$

If the imputation process is performed independently for each missing y_i, the imputation variance of $\hat{Y}_{kI}$ is given by

$$E_p E_q V_I\left(\hat{Y}_{kI} \middle| s\right) = E_p E_q \left[\frac{1}{N^2}\sum_{i\in s} w_i^2 (1-r_i)\hat{\varphi}_i^k (1-\hat{\varphi}_i^k)\right].$$

The balanced random imputation method consists of selecting the imputed values y_i^* so that the following equations are approximately satisfied

$$\sum_{i\in s} w_i (1-r_i) 1(y_i^* = k) = \sum_{i\in s} w_i (1-r_i)\hat{\varphi}_i^k \quad \text{for any} \quad k = 1, \cdots, K. \tag{5.5}$$

Now consider the $n_m \times K$ table

	1	…	k	…	K
1	$(\psi_{11}, \mathbf{e}_{11})$	…	$(\psi_{1k}, \mathbf{e}_{1k})$	…	$(\psi_{1K}, \mathbf{e}_{1K})$
⋮	⋮		⋮		⋮
i	$(\psi_{i1}, \mathbf{e}_{i1})$	…	$(\psi_{ik}, \mathbf{e}_{ik})$	…	$(\psi_{iK}, \mathbf{e}_{iK})$
⋮	⋮		⋮		⋮
n_m	$(\psi_{n_m 1}, \mathbf{e}_{n_m 1})$	…	$(\psi_{n_m k}, \mathbf{e}_{n_m k})$	…	$(\psi_{n_m K}, \mathbf{e}_{n_m K})$

where each cell (i,k) is given the vector value $\mathbf{e}_{ik}$ which is the column vector of size K with 1 on row k and 0 elsewhere, and the probability $\psi_{ik} = \hat{\varphi}_i^k$. A balanced random imputation such that equations (5.5) are approximately respected may be handled with the Cube method presented in the Appendix, with inclusion probabilities ψ_{ik} and the vector $\tilde{\mathbf{x}}_{ik} = \left((\mathbf{x}_{ik}^0)', x_{ik}^1, \cdots, x_{ik}^{n_m}\right)'$, where $\mathbf{x}_{ik}^0 = w_i \psi_{ik} \mathbf{e}_{ik}$ and $x_{ik}^{i^*} = \psi_{ik} \delta_{i^* i}$ for $i^* = 1, \cdots, n_m$, where $\delta_{i^* i} = 1$ if $i^* = i$ and $\delta_{i^* i} = 0$, otherwise.

6. SIMULATION STUDY

We conducted a simulation study to investigate on the performance of the proposed balanced random imputation method. We used a population of size $N = 10,000$ consisting of a data file that was extracted from a sample collected between January 2005 and December 2005 for the Canadian Community Health Survey (CCHS), which is a cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population. We considered two variables of interest: the self reported weight in kilograms of an individual (y_1), which is a continuous variable, and the self reported

presence of asthma (y_2) such that $y_2 = 1$ if the individual reported suffering from asthma and $y_2 = 0,$ otherwise. The population was stratified into the Canadian provinces, which led to the creation of 11 strata.

For the variable $y_1,$ our objective was to estimate two parameters: (i) its population mean, $\bar{Y}_1 = \frac{1}{N}\sum_{i \in U} y_{1i}$ and (ii) its finite population distribution function, $F_{1N}(t) = \frac{1}{N}\sum_{i \in U} \mathbf{1}(y_{1i} \le t),$ for different values of t (0.05; 0.25; 0.5; 0.75; 0.95). For the variable $y_2,$ our objective was to estimate the proportion of individuals in the population that reported suffering from asthma, $\bar{Y}_2 = \frac{1}{N}\sum_{i \in U} y_{2i}.$

From the population, we selected 1000 stratified simple random samples without replacement of size n = 500 using proportional allocation. That is, the sample size n_h in stratum h was set to $n_h = 500\frac{N_h}{N},$ where N_h denotes the number of individual in stratum $h,$ $h = 1,...,8.$ From each generated sample, we generated nonresponse to the variables y_1 and y_2 according to a uniform response mechanism within strata. That is, the probability of response within strata $h,$ $p_h,$ is constant but varies across strata. Also, units within a stratum respond independently from one another. Table 6.1 shows useful quantities for the CCHS population.

The response indicators r_i for $i \in s,$ were then generated independently 1000 times from a Bernoulli distribution with parameter $p_h,$ $h = 1,...,11,$ which led to 1000 sets of respondents. In each sample containing respondents and nonrespondents, imputation was performed within each stratum independently. In other words, the strata were used as imputation classes. Within

each class, we performed the imputations according to three methods: (i) mean imputation within classes (MI), (ii) random hot-deck within classes (RHDI) and (iii) random balanced imputation (RBALI). Note that RHDI can be viewed as MI with added residuals. The two latter methods are described in section 2.

Table 6.1: Parameters used for the CCHS population

Strata	1	2	3	4	5	6	7	8	9	10	11
N_h	641	433	693	689	1435	1919	794	845	992	1096	463
n_h	32	22	35	34	72	96	40	42	50	55	22
p_h	0.7	0.7	0.7	0.7	0.7	0.7	0.6	0.6	0.6	0.6	0.6

As a measure of the bias of an estimator $\hat{\theta}$ of a parameter θ, we used the Monte Carlo Percent Relative Bias (RB) given by

$$RB\left(\hat{\theta}\right) = \frac{E_{MC}\left(\hat{\theta}\right) - \theta}{\theta} \times 100, \tag{6.1}$$

where $E_{MC}\left(\hat{\theta}\right) = \frac{1}{1000}\sum_{r=1}^{1000}\hat{\theta}^{(r)}$, and $\hat{\theta}^{(r)}$ denotes the estimator $\hat{\theta}$ in the r-th sample, $r = 1,...,1000$. As a measure of variability of $\hat{\theta}$, we used the Monte Carlo Mean Square Error (MSE) given by

$$MSE_{MC}\left(\hat{\theta}\right) = \frac{1}{1000}\sum_{r=1}^{1000}\left(\hat{\theta}^{(r)} - \theta\right)^2. \tag{6.2}$$

Let $\hat{\theta}^{(\mathrm{MI})}, \hat{\theta}^{(\mathrm{RHDI})}$ and $\hat{\theta}^{(\mathrm{RBALI})}$ denote the estimator $\hat{\theta}$ under MI, RHDI and RBALI. In order to compare the relative stability of the imputed estimators, using $\hat{\theta}^{(\mathrm{RHDI})}$ as the reference, we used the following measure:

$$RE = \frac{MSE_{MC}\left(\hat{\theta}^{(\cdot)}\right)}{MSE_{MC}\left(\hat{\theta}^{(\text{RHDI})}\right)}. \tag{6.3}$$

When $\theta = \bar{Y}_1$, we have $\hat{\theta} \equiv \hat{\bar{Y}}_{1I} = \frac{\hat{Y}_{1I}}{N}$, where $\hat{Y}_{1I}$ is obtained from (2.2) by replacing y_i with y_{1i} for $i \in s_r$ and y_i^* with y_{1i}^* for $i \in s_m$. When $\theta = F_{1N}(t)$, we have $\hat{\theta} \equiv \hat{F}_{1I}(t)$, where $\hat{F}_{1I}(t)$ is obtained from (4.1) by replacing y_i with y_{1i} for $i \in s_r$ and y_i^* with y_{1i}^* for $i \in s_m$. Finally, when $\theta = \bar{Y}_2$, we have $\hat{\theta} \equiv \hat{\bar{Y}}_{2I} = \frac{\hat{Y}_{2I}}{N}$, where $\hat{Y}_{2I}$ is obtained from (2.2) by replacing y_i with y_{2i} for $i \in s_r$ and y_i^* with y_{2i}^* for $i \in s_m$.

Table 6.3 shows the monte carlo percent relative bias (RB) of the imputed estimator and the RE, which are obtained from (6.1)-(6.3) by replacing $\hat{\theta}$ with $\hat{\bar{Y}}_{1I}$ and θ with $\bar{Y}_1$. First, the imputed estimator is approximately unbiased in all the scenarios, as expected. In terms of RE, results show that $\hat{\bar{Y}}_{1I}^{(\text{MI})}$ has the smallest MSE. This result is not surprising since the imputation variance is identically equal to zero in this case. Also, it is clear that $\hat{\bar{Y}}_{1I}^{(\text{RBALI})}$ is significantly more efficient than $\hat{\bar{Y}}_{1I}^{(\text{RDHI})}$ with a value of RE equal to 0.82. Finally, $\hat{\bar{Y}}_{1I}^{(\text{RBALI})}$ is slightly less efficient than $\hat{\bar{Y}}_{1I}^{(\text{MI})}$. This is due to the fact that the balancing equations needed to be relaxed in the last phase of the imputation algorithm in order to end the selection of residuals for non-responding units. As a result, equation (2.10) did not to hold exactly and so the imputation variance was not entirely eliminated.

Table 6.3: Monte Carlo percent Relative Bias of the imputed estimator and RE

	MI	RHDI	RBALI

CCHS population	$RB(\hat{\bar{Y}}_{1I})$	0.01	0.02	0.01
	RE	0.80	1	0.82

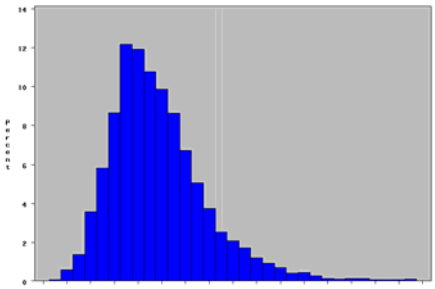
We now turn to the distribution function, $F_{1N}(t)$. Table 6.4 shows the RB of the imputed estimator and the RE, which are obtained from (6.1)-(6.3) by replacing $\hat{\theta}$ with $\hat{F}_{1I}(t)$ and θ with $F_{1N}(t)$. As expected, the distribution function under MI is significantly biased. In terms of relative bias, both RBALI and RHDI show almost not bias (less than 0.25%), as expected. Both imputation methods succeed in preserving the distribution of the variable y_1. Also, it is clear that the imputed estimator under RBALI is more efficient then the corresponding estimator under RHDI in all the scenarios. Figure 3 shows that the distribution function is preserved under RHDI and RBALI, unlike MI which leads to a considerable distortion of the population distribution function. In the latte case, we note the occurrence of a spike at the respondent mean.

Table 6.4: Monte Carlo percent Relative Bias of the imputed estimator of the distribution function and RE

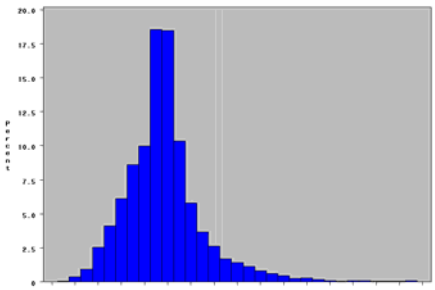
$F(t)$	*RB*			*RE*	
	RHDI	*RHDI*	*RBALI*	*RBALI*	*RBALI*
.05	-29.05	0.05	0.05	1.88	0.99
.25	-29.03	0.18	0.25	9.52	0.90
.50	-16.30	-0.11	-0.02	11.02	0.90
.75	8.94	0.03	0.01	7.90	0.87
.95	1.57	-0.06	-0.05	1.78	0.91

Figure 3: Population Distribution Function (PDF) and Monte Carlo Mean of the Estimated Distribution Functions (EDF)

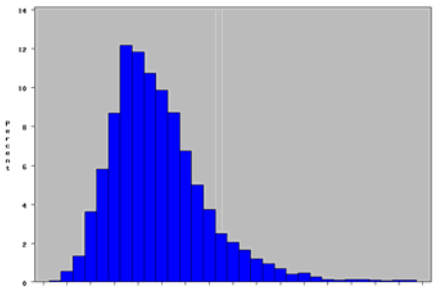
PDF



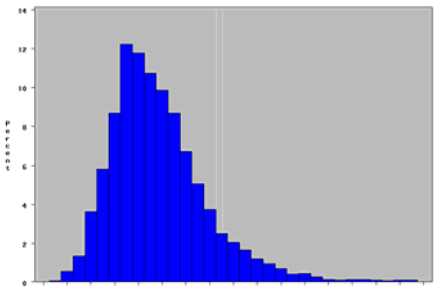
EDF under MI



EDF under RHDI



EDF under RBALI



Finally, we turn to the binary variable y_2. Table 7.5 shows the RB of the imputed estimator and the RE, which are obtained from (6.1)-(6.3) by replacing $\hat{\theta}$ with $\hat{\bar{Y}}_{2I}$ and θ with $\bar{Y}_2$. Here, the results are very similar to those obtained for the variable y_1….. Note that in practice, MI in this context is seldom used because it leads to the creation of impossible values in the data file.

Table 6.5: Monte Carlo percent Relative Bias of the imputed estimator and RE

	MI	RHDI	RBALI
$RB\left(\hat{\bar{Y}}_{2I}\right)$	-0.03	-0.01	-0.02
RE	0.82	1	0.86

7. SUMMARY AND DISCUSSION

In this paper, we have studied the problem of balanced random imputation as a way to reduce/eliminate the imputation variance, which is often viewed as a parasitic variance. We proposed a general algorithm for selecting the random residuals that was inspired from the Cube method proposed by Deville and Tillé (2004) in the context of balanced sampling. The proposed algorithm can be applied for both continuous and categorical variables and for any sampling design and imputation method. Results from a limited simulation study have shown that in all the cases the proposed balanced random imputation method was efficient in comparison to the corresponding random imputation method.

If that the balancing constraints are exactly satisfied, the variance of the imputed estimator (2.2) can be readily estimated using any variance estimation available; e.g., Rao and Shao (1992), Särndal (1992) and Shao and Steel (1999). If the balancing equations are not exactly satisified, the Cube method involves a rounding process called the landing phase. In this case, correct variance estimation is not straightforward because it involves estimating the variance due to the landing phase. This problem is currently under investigation.

In practice, estimates of bivariate parameters such as domain means, regression coefficients and coefficients of correlation are often needed. In this case, determining an imputation method that preserves the relationships between variables becomes the main challenge. The use of balanced imputation to overcome this problem is currently under investigation.

REFERENCES

Chen, H.L., Rao, J. N. K. and Sitter, R. R. (2000). Efficient Random Imputation for Missing Survey Data in Complex Survey. *Statistica Sinica*, 10, pp. 1153-1169.

Deville, J-C. (2006). Random Imputation Using Balanced Sampling. Presentation to the JointStatistical Meeting of the American Statistical Association, Seattle, USA.

Deville, J-C., and Tillé, Y. (2004). Efficient balanced sampling : the Cube method. Biometrika, 91, pp. 893-912.

Deville, J-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. Journal of Statistical Planning and inference, 128, pp. 569-591.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. Journal of the American Statistical Association, 91, pp. 490-498.

Fuller, W.A. and Kim, J.K. (2005). Hot-deck imputation for the response model. Survey Methodology, 31, pp. 139-149.

Hajek (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 35, pp. 1491-1523.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under a regression superpopulation model. Journal of the American Statistical Association, 77, pp. 89-96.

Haziza, D. (2009), Imputation and inference in the presence of missing data. To appear in Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference, Editors: C.R. Rao and D. Pfeffermann.

Kalton, G. and Kish, L. (1981). Two efficient random imputation procedures. *Proceedings of the Survey Research Methods, American Statistical Association*, pp. 146-151.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. Communications in Statistics, Part A- Theory and Methods, 13, pp. 1919-1939.

Kim, J.K. and Fuller, W.A. (2004), Fractional hot-deck imputation. Biometrika, 91, pp. 559-578.

Rao, J. N. K. and Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811-822.

Rubin, D. B. (1987). Multiple imputations for nonresponse in surveys. Wiley, New York.

Särndal, C. E. (1992), "Method for estimating the precision of survey estimates when imputation has been used", *Survey Methodology*, 18, pp. 241-252.

Shao, J. and Steel, P. (1999), "Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions", *Journal of the American Statistical Association*, 94, pp. 254-265.

APPENDIX : THE CUBE METHOD

Before introducing the different steps of the algorithm, we introduce further notation. The population U^* is partitioned into n_m strata $U_1^*,\cdots,U_i^*,\cdots U_{n_m}^*$ 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 p_{ij} and $\tilde{\mathbf{x}}_{ij}$ (see section 3). Let $A=\left(\frac{\tilde{\mathbf{x}}_{11}}{p_{11}},\cdots,\frac{\tilde{\mathbf{x}}_{ij}}{p_{ij}},\cdots,\frac{\tilde{\mathbf{x}}_{n_m n_r}}{p_{n_m n_r}}\right)$ be a $n_m \times n_r$ matrix called the matrix of constraints. Then the first phase of the imputation algorithm follows these steps. We first initialize with $\phi(0)=\phi.$ Next, at time $t=1,\cdots,T$, repeat the following three steps :

Step 1 : Generate any vector $u(t)=\left(u_{11}(t),\cdots,u_{ij}(t),\cdots,u_{n_m n_r}(t)\right)' \neq 0$ such that

(1) $u(t)$ is in the kernel of the matrix A

(2) $u_{ij}(t)=0$ if $\phi_{ij}(t-1)$ is an integer.

Step 2 : Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$ the largest values of $\lambda_1(t)$ and $\lambda_2(t)$ such that

$$0 \le \phi(t-1) + \lambda_1(t)u(t) \le 1,$$

$$0 \le \phi(t-1) - \lambda_2(t)u(t) \le 1.$$

Note that $\lambda_1^*(t) > 0$ and $\lambda_2^*(t) > 0$.

Step 3 : Select

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

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

The choice of $u(t)$ in Step 1 implies that at each step of the former algorithm, the balancing equations remain exactly respected. The choice of $\lambda_1^*(t)$ and $\lambda_2^*(t)$ in Step 2 imply that the vector $\phi(t)$ has one more integer component than $\phi(t-1)$. This means that at each step, one more unit is either sampled or definitely rejected. Finally, the random choice in Step 3 implies that the inclusion probabilities are also exactly respected. The algorithm stops when it is no more possible to select a vector $u(t)$ such that (1) and (2) are satisfied. T denotes the time when the flight phase stops. Let $\phi^* = \phi(T)$, and r denote the number of non integer components in ϕ^*. Theorem 8.1 in Tillé (2001, p.??) implies that r is no greater than the number of balancing variables, which equals $n_m + 1$. We show below that r is in fact no greater than 2.

First, suppose that there exists a stratum U_i^* in which the vector ϕ^* has at least 3 non-integer components. We assume without loss of generality that this stratum is U_1^*, and that ϕ_{11}^*, ϕ_{12}^* and ϕ_{13}^* are not integer. Let

$$A_1 = \begin{pmatrix} \frac{x_{11}^0}{p_{11}} & \frac{x_{12}^0}{p_{12}} & \frac{x_{13}^0}{p_{13}} \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ & & \end{pmatrix}$$

be the square sub-matrix given by the first three lines and the first three columns in A. A_1 is obviously singular, and if we denote $v_3 = (v_{11}, v_{12}, v_{13})'$ for a vector in the kernel of A_1, v_3 may be filled with zeros to form a vector v in the kernel of A such that $v_{ij}(t) = 0$ if $\phi_{ij}(T)$ is an integer, which is impossible. Consequently, each stratum U_i^* has at most two non-integer components.

Now, suppose that there exists at least two strata U_i^* and U_k^* in which the vector ϕ^* has non-integer components. We assume without loss of generality that such strata are given by U_1^* and U_2^*, and that ϕ_{11}^*, ϕ_{12}^*, ϕ_{21}^*, ϕ_{22}^* are not integer. Let

$$A_2 = \begin{pmatrix} \frac{x_{11}^0}{p_{11}} & \frac{x_{12}^0}{p_{12}} & \frac{x_{21}^0}{p_{21}} & \frac{x_{22}^0}{p_{22}} \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

be the square sub-matrix given by the first four lines and columns 1, 2, $n_m + 1$ and $n_m + 2$, in A. A_2 is obviously singular, and if we denote $v_4 = (v_{11}, v_{12}, v_{21}, v_{22})'$ for a vector in the kernel of A_2, v_4 may be filled with zeros to form a vector w in the kernel of A such that $v_{ij}(t) = 0$ if $\phi_{ij}(T)$ is an integer, which is impossible.

Consequently, there is at most one stratum U_i^* which has non-integer components, and this stratum has no more than two non-integer components. At the end of the first phase, this is no

more possible for the balancing on variable x^0 to hold exactly. In a second phase, this condition is suppressed, and a random choice is made between the two (possibly) remaining units for the inclusion probabilities to remain exactly respected and the condition of fixed size in each stratum to hold exactly.