

On Balanced Random Imputation in Surveys

Guillaume Chauvet, Jean-Claude Deville and David Haziza

Summary

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. A special case of random imputation, random hot-deck imputation, is often used in practice if the variable being imputed is categorical because it eliminates the possibility of impossible values. Also, it is used when it is desired to impute more than one variable at the time because the same donor can be used to impute all the missing values, which helps preserving the relationships between variables. 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. The proposed class of imputation methods can be applied for both categorical and continuous variables. The results of a limited simulation study support our findings.

Keywords : Balanced sampling ; deterministic imputation ; distribution function ; imputation variance ; random imputation.

1 Introduction

To compensate for item nonresponse in surveys, some form of imputation is typically used. 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. 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. Random hot-deck imputation is often used in practice if the variable being imputed is categorical because it eliminates the possibility of impossible values in the imputed data file. Also, it is used when it is desired to impute more than one variable at the time because the same donor can be used to impute all the missing values, which helps preserving the relationships between variables.

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 of totals if the underlying imputation or nonresponse model is correctly specified (e.g., Haziza, 2009), 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.

One important 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. 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 reduce (or eliminate) 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 since imputing multiple values is often seen as operationally less convenient than single imputation. 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 methods 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) in the context of balanced sampling. 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 eliminating (or at least, significantly reducing) the imputation variance significantly while preserving the distribution of the variable being imputed. In our view, the proposed method should be preferred even as compared to other methods of the third approach, since it enables a random selection of donors which is perfectly adapted to the parameter of interest. That is, the donors are randomly selected for the imputation variance to be eliminated, and not only to be reduced. Also, additional constraints may be easily included in the imputation mechanism if other functions of the imputed variable are of interest (see section 4), since the Cube method may handle an arbitrary number of balancing variables.

In our view, the third approach is more attractive than the first two 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 that lead to estimators of totals almost as efficient than those that would have been obtained under a given deterministic imputation method and at the same time, lead to asymptotically unbiased estimators of quantiles.

2 Notation and Random Imputation

Let $U = \{1, 2, \dots, N\}$ be a finite population consisting of N elements. We consider the problem of estimating the census regression coefficient,

$$\theta_N = \left(\sum_{i \in U} x_i x_i^\top \right)^{-1} \sum_{i \in U} x_i y_i,$$

where y_i denotes the i -th value of the variable of interest y , and x_i is a l -vector of variables attached to unit i , $i = 1, \dots, N$. When $x_i = 1$ for all $i \in U$, θ_N reduces to the overall population mean of the y -values, $\bar{Y}_N = N^{-1} \sum_{i \in U} y_i$. Another special case of θ_N occurs when x_i is a vector of domain indicators, in which case $\theta_N = (\bar{Y}_1, \dots, \bar{Y}_l)^\top$ is the vector of domain means.

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 $d_i = 1/\pi_i$ denote its design weight. In the absence of nonresponse, a basic

estimator $\hat{\theta}_n$ of θ_N is given by

$$\hat{\theta}_n = \left(\sum_{i \in s} d_i x_i x_i^\top \right)^{-1} \sum_{i \in s} d_i x_i y_i. \quad (1)$$

The estimator $\hat{\theta}_n$ in (1) is asymptotically design-unbiased (or p -unbiased) for θ_N .

In this paper, we assume that the vector x_i is observed for all $i \in s$, whereas the variable y is potentially missing. Let y_i^* denote the imputed value used to replace the missing y_i . An imputed estimator of θ_N is given by

$$\hat{\theta}_I = \hat{T}^{-1} \left(\sum_{i \in s} d_i r_i x_i y_i + \sum_{i \in s} d_i (1 - r_i) x_i y_i^* \right), \quad (2)$$

where $\hat{T} = (\sum_{i \in s} d_i x_i x_i^\top)$ and 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 nonrespondents of size n_m .

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

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

where $f(\cdot)$ is a given function, $z = (z_1, \dots, z_K)^\top$ is a K -vector of auxiliary variables available at the imputation stage for all the sampled units, β is a K -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, and their common distribution function is denoted by $F_\epsilon(\cdot)$. The model (3) is often called an imputation model (e.g., Särndal, 1992). In the case of deterministic imputation, the imputed value y_i^* is given by $y_i^* = f(z_i; \hat{B}_r)$, for $i \in s_m$, where \hat{B}_r is an estimator of β , which is solution of the

following estimating equations :

$$\sum_{i \in s} \omega_i r_i v_i^{-1} [y_i - f(z_i; \beta)] h_i = 0, \quad (4)$$

where $h_i = \partial f(z_i; \beta) / \partial \beta$ and ω_i is an imputation weight attached to unit i . Several choices of ω_i are possible : the choice $\omega_i = d_i$ leads to the customary survey weighted imputation, whereas the choice $\omega_i = 1$ leads to unweighted imputation. Other choices of imputation weights are possible ; e.g., Haziza (2009).

Random imputation can be seen as a deterministic imputation to which a random noise ϵ_i^* is added. That is, the imputed value y_i^* is given by

$$y_i^* = f(z_i; \hat{B}_r) + \hat{\sigma} \sqrt{v_i} \epsilon_i^* \text{ for } i \in s_m, \quad (5)$$

where $\hat{\sigma}$ is an estimator of σ . The random quantity ϵ_i^* can be generated from a given parametric distribution. However, in practice, it is natural to generate these random quantities from the empirical distribution function of the residuals. More precisely, we assume that the random residuals ϵ_i^* are selected independently and with replacement from the set, $E_r = \{\tilde{e}_j; j \in s_r\}$, of standardized residuals observed from the responding units, with probabilities

$$pr(\epsilon_i^* = \tilde{e}_j) = \omega_j / \sum_{l \in s} \omega_l r_l, \quad (6)$$

where $\tilde{e}_j = e_j - \bar{e}_r$, $e_j = \hat{\sigma}^{-1} v_j^{-1/2} \{y_j - f(z_j; \hat{B}_r)\}$ with $\bar{e}_r = \sum_{j \in s} \omega_j r_j e_j / \sum_{j \in s} \omega_j r_j$. 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,

$$\hat{F}_{\epsilon, r}(t) = \sum_{j \in s} \tilde{\omega}_j r_j 1(\tilde{e}_j \leq t), \quad (7)$$

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

Note that the imputed value y_i^* in (5) can be viewed as the sum of a deterministic component, $f(z_i; \hat{B}_r)$, and a random component ϵ_i^* . In other words, for each random imputation, there is a corresponding deterministic imputation, which is obtained by setting $\epsilon_i^* = 0$ for all i . Letting $f(z_i; \beta) = z_i^\top \beta$ in (3) leads to the model underlying (deterministic and random) regression imputation. Random regression imputation is thus obtained from (5) by setting $f(z_i; \hat{B}_r) = z_i^\top \hat{B}_r$, where

$$\hat{B}_r = \left[\sum_{i \in s} \omega_i r_i z_i z_i^\top / v_i \right]^{-1} \sum_{i \in s} \omega_i r_i z_i y_i / v_i$$

is a solution of the estimating equations (4) with $f(z_i; \beta) = z_i^\top \beta$ and $h_i = z_i$. That is, random regression imputation uses the imputed values

$$y_i^* = z_i^\top \hat{B}_r + \hat{\sigma} \sqrt{v_i} \epsilon_i^*. \quad (8)$$

Random hot-deck imputation within classes, which is a popular method in practice, can be viewed as a special case of (8). 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; random hot-deck imputation within classes is obtained from (8) by setting $z_i = (z_{1i}, \dots, z_{Ki})^\top$ and $v_i = v_k$ if i belongs to class k .

3 Balanced Random Imputation

In this section, we discuss the concept of balanced imputation. Using the imputed values (5) in (2) leads to

$$\hat{\theta}_I = \hat{T}^{-1} \left(\sum_{i \in s} d_i r_i x_i y_i + \sum_{i \in s} d_i (1 - r_i) x_i f(z_i; \hat{B}_r) + \sum_{i \in s} d_i (1 - r_i) x_i \hat{\sigma} \sqrt{v_i} \epsilon_i^* \right). \quad (9)$$

The total error of $\hat{\theta}_I$ can be expressed as

$$\hat{\theta}_I - \theta_N = \left(\hat{\theta}_n - \theta_N \right) + \left(E_I(\hat{\theta}_I) - \hat{\theta}_n \right) + \left(\hat{\theta}_I - E_I(\hat{\theta}_I) \right), \quad (10)$$

where $E_I(\cdot)$ denotes the expectation with respect to the imputation mechanism, which is used to select the random residuals. The first term on the right hand side of (10) represents the sampling error, whereas the second and the third terms represent the nonresponse error and the imputation error, respectively. When taking expectations with respect to the imputation mechanism, note that only the ϵ_i^* 's are treated as random, whereas all the other variables are treated as fixed. Note that the imputation error, $\hat{\theta}_I - E_I(\hat{\theta}_I)$, is equal to 0 in the case of deterministic imputation. Let the subscripts p and q indicate the sampling mechanism and the nonresponse mechanism, respectively. The total variance of $\hat{\theta}_I$ given by (2) can be expressed as

$$V(\hat{\theta}_I) = V_p E_q E_I(\hat{\theta}_I) + E_p V_q E_I(\hat{\theta}_I) + E_p E_q V_I(\hat{\theta}_I), \quad (11)$$

where $V_I(\cdot)$ denotes the variance with respect to the imputation mechanism. The first term on the right hand side of (11) is the sampling variance, the second term is the nonresponse variance, whereas the third term is the imputation variance. The magnitude of both the nonresponse and imputation variances typically depends on the response rate and the predictive power of the imputation model.

We focus on the imputation variance, which is given by

$$E_p E_q V_I(\hat{\theta}_I) = E_p E_q \left[\hat{T}^{-1} \left(\sum_{i \in s} \tilde{x}_i \tilde{x}_i^\top \right) \frac{\sum_{i \in s} \omega_i r_i \tilde{\epsilon}_i^2}{\sum_{i \in s} \omega_i r_i} \hat{T} \right], \quad (12)$$

where $\tilde{x}_i = d_i(1 - r_i)x_i\sqrt{v_i}$. Under mild regularity conditions, the imputation variance given by (12) is $O(1/n)$ which is the same order of magnitude as the sampling and nonresponse variances. From (12), we note that the imputation variance is small if (i) the response rate is high, in which case the term $\sum_{i \in s} \tilde{x}_i \tilde{x}_i^\top$ is likely to be small (in fact, in the complete data case, we have $r_i = 1$ for all i and so this term vanishes) or if (ii) the residuals $\tilde{\epsilon}_i$ 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. In other words, the imputation variance is essentially a parasitic variance that should be eliminated or, at least, significantly reduced.

In order to eliminate the imputation variance of $\hat{\theta}_I$, the random residuals must be selected in such a way that the imputation error

$$\hat{\theta}_I - E_I(\hat{\theta}_I) = \sum_{i \in s} d_i(1 - r_i)x_i\hat{\sigma}\sqrt{v_i}\epsilon_i^*$$

is eliminated. We thus propose a balanced random imputation method which consists of selecting the residuals so that the following equation is (approximately) satisfied :

$$\sum_{i \in s} d_i(1 - r_i)x_i\hat{\sigma}\sqrt{v_i}\epsilon_i^* = 0. \quad (13)$$

More precisely, the imputed values under random balanced imputation are given by

$$y_i^{**} = f(z_i; \hat{B}_r) + \hat{\sigma}\sqrt{v_i}\epsilon_i^{**}, \quad (14)$$

for $i \in s_m$, where the random residuals ϵ_i^{**} are selected from the set, $E_r =$

$\{\tilde{\epsilon}_j; j \in s_r\}$, with probabilities

$$pr(\epsilon_i^{**} = \tilde{\epsilon}_j) = \omega_j / \sum_{l \in s} \omega_l r_l, \quad (15)$$

so that the equation (13) (approximately) holds. This imputation procedure can be performed by adapting the random Cube algorithm originally proposed by Deville and Tillé (2004) in the context of balanced sampling. The algorithm for balanced imputation is described in details in Chauvet, Deville and Haziza (2010). If the equation (13) is exactly satisfied, then the imputation variance is completely eliminated and the resulting estimator is fully efficient (e.g., Kim and Fuller, 2004). In some situations, it is not possible to satisfy (13) exactly but only approximately. In this case, the imputation variance is not completely eliminated but it is expected to be significantly reduced. In section 4, we show that the empirical distribution function based on observed values and imputed values given by (14) is a consistent estimator of the true population distribution function. In other words, the proposed balanced imputation method preserves the distribution of the variable being imputed. Balanced random regression imputation is obtained from (14) by setting $f(z_i; \hat{B}_r) = z_i^\top \hat{B}_r$. That is, balanced random regression imputation uses the imputed values

$$y_i^{**} = z_i^\top \hat{B}_r + \hat{\sigma} \sqrt{v_i} \epsilon_i^{**}. \quad (16)$$

We make the following remarks : first, suppose that $\theta_N = \bar{Y}_N$, (which occurs when $x_i = 1$ for all $i \in U$) and that random hot-deck within classes is used to compensate nonresponse to variable y . Then, the condition (13) reduces to

$$\bar{y}_{mk}^{**} = \bar{y}_{rk}, \quad k = 1, \dots, K, \quad (17)$$

where

$$\bar{y}_{rk} = \frac{\sum_{i \in s_k} \omega_i r_i y_i}{\sum_{i \in s_k} \omega_i r_i}$$

is the weighted mean of the respondents in class k and

$$\bar{y}_{mk}^{**} = \frac{\sum_{i \in s_k} \omega_i (1 - r_i) y_i^{**}}{\sum_{i \in s_k} \omega_i (1 - r_i)}$$

is the weighted mean of the imputed values in class k . In other words, eliminating the imputation variance will consist of 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 random hot-deck imputation so that (17) 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 (17) is satisfied, which is more attractive from a data user's perspective.

Secondly, suppose we are interested in estimating the vector of domain means, $\theta_N = (\bar{Y}_1, \dots, \bar{Y}_l)^\top$. Satisfying (13) requires the domains to be specified at the imputation stage, which may not be the case in practice. Domains are sometimes specified at the estimation stage, after the imputation process has been completed. In this case, the resulting imputed domain estimators may be substantially biased if these domains are related to the variable being imputed. A possible solution is to determine a bias-adjusted estimator; e.g., Haziza and Rao (2005). This situation is not considered in this paper.

Finally, let $\hat{\beta}_I$ be an imputed estimator of β defined as

$$\hat{\beta}_I = \left(\sum_{i \in s} d_i v_i^{-1} z_i z_i^\top \right)^{-1} \left[\sum_{i \in s} d_i r_i v_i^{-1} z_i y_i + \sum_{i \in s} d_i (1 - r_i) v_i^{-1} z_i y_i^* \right]. \quad (18)$$

As pointed out by a referee, adding the constraint

$$\sum_{i \in s} d_i (1 - r_i) v_i^{-1} z_i [\hat{\sigma} \sqrt{v_i} \epsilon_i^*] = 0 \quad (19)$$

will make $\hat{\beta}_I$ fully efficient for β . That is, $V_I(\hat{\beta}_I) = 0$. Thus, satisfying (13) and (19) will make both $\hat{\theta}_I$ and $\hat{\beta}_I$ fully efficient.

4 Estimation of other functions

In sections 2 and 3, we considered the case of a census regression coefficient, which is linear in y since it can be expressed as

$$\theta_N = \sum_{i \in U} a_i y_i,$$

where $a_i = (\sum_{i \in U} x_i x_i^\top)^{-1} x_i$. We now turn to the case of the following finite population parameter :

$$\theta_N = \sum_{i \in U} \phi(y_i), \quad (20)$$

where $\phi(\cdot)$ is a differentiable function with continuous partial derivatives of order 2. For simplicity, we consider the case of scalar θ_N . For example, if $\phi(y_i) = y_i^2$, then $\theta_N = \sum_{i \in U} y_i^2$, the sum of square of the y -values. A complete data estimator of θ_N is now given by $\hat{\theta}_n = \sum_{i \in s} d_i \phi(y_i)$. In the presence of nonresponse to item y , an imputed estimator of θ_N is defined as

$$\hat{\theta}_I = \sum_{i \in s} d_i r_i \phi(y_i) + \sum_{i \in s} d_i (1 - r_i) \phi(y_i^*), \quad (21)$$

where y_i^* is given by (5). The imputation error of $\hat{\theta}_I$ given in (21), $\hat{\theta}_I - E_I(\hat{\theta}_I)$, is given by

$$\hat{\theta}_I - E_I(\hat{\theta}_I) = \sum_{i \in s} d_i (1 - r_i) [\phi(y_i^*) - E_I(\phi(y_i^*))]. \quad (22)$$

It follows from (22) that the imputation variance of $\hat{\theta}_I$, $V_I(\hat{\theta}_I)$, vanishes if

$$\sum_{i \in s} d_i(1 - r_i) [\phi(y_i^*) - E_I(\phi(y_i^*))] = 0. \quad (23)$$

Assuming that the function ϕ is $m + 1$ times differentiable and that the derivatives are continuous, we have

$$\begin{aligned} \phi(y_i^*) &= \phi(\hat{y}_i + \hat{\sigma}v_i^{-1/2}\epsilon_i^*) \\ &= \sum_{k=0}^m \frac{\phi^{(k)}(\hat{y}_i)}{k!} (\hat{\sigma}v_i^{-1/2}\epsilon_i^*)^k + \frac{\phi^{(m+1)}(\alpha_i)}{(m+1)!} (\hat{\sigma}v_i^{-1/2}\epsilon_i^*)^{m+1}, \end{aligned} \quad (24)$$

where $\alpha_i \in (y_i^*, \hat{y}_i)$. It follows that

$$\begin{aligned} &\hat{\theta}_I - E_I(\hat{\theta}_I) \\ &= \sum_{i \in s} d_i(1 - r_i) \left\{ \sum_{k=0}^m \frac{\phi^{(k)}(\hat{y}_i)}{k!} (\hat{\sigma}v_i^{-1/2}\epsilon_i^*)^k + \xi_{i,m+1} \right\} \\ &- \sum_{i \in s} d_i(1 - r_i) \left\{ \sum_{k=0}^m \frac{\phi^{(k)}(\hat{y}_i)}{k!} E_I \left[(\hat{\sigma}v_i^{-1/2}\epsilon_i^*)^k \right] + E_I(\xi_{i,m+1}) \right\}, \end{aligned} \quad (25)$$

where

$$\xi_{i,m+1} = \frac{\phi^{(m+1)}(\alpha_i)}{(m+1)!} (\hat{\sigma}v_i^{-1/2}\epsilon_i^*)^{m+1}.$$

We can rewrite (25) as

$$\begin{aligned} &\hat{\theta}_I - E_I(\hat{\theta}_I) \\ &= \sum_{k=0}^m \left[\sum_{i \in s} d_i(1 - r_i) \frac{\phi^{(k)}(\hat{y}_i)}{k!} \left\{ (\hat{\sigma}v_i^{-1/2}\epsilon_i^*)^k - E_I \left[(\hat{\sigma}v_i^{-1/2}\epsilon_i^*)^k \right] \right\} \right] \\ &+ \sum_{i \in s} d_i(1 - r_i) [\xi_{i,m+1} - E_I(\xi_{i,m+1})]. \end{aligned} \quad (26)$$

The first term on the right hand side of (26) defines a system of m balancing equations. If the random residuals are selected in such a way that this term is zero, the residual imputation error comes from the term $m + 1$ or higher. Consequently, we may expect the imputation variance of $\hat{\theta}_I$ to be significantly reduced if the corresponding balancing equations are used in the imputation mechanism.

5 Consistency of the distribution function

In this section, we study the asymptotic properties of the estimated distribution function under the imputation mechanisms (8) and (16), respectively. We assume that there is a sequence of sampling designs and finite populations, indexed by ν , such that the population size N_ν , the sample size n_ν and the number of respondents n_{r_ν} tend to infinity as $\nu \rightarrow \infty$. Though the subscript ν is suppressed for convenience, all the limiting processes are understood to be as $\nu \rightarrow \infty$.

The finite population distribution function of the variable y can be written as

$$F_N(t) = \frac{1}{N} \sum_{i \in U} 1(y_i \leq t). \quad (27)$$

A complete data estimator of $F_N(t)$ is given by

$$\hat{F}_N(t) = \sum_{i \in S} \tilde{d}_i 1(y_i \leq t), \quad (28)$$

where $\tilde{d}_i = d_i / \sum_{j \in S} d_j$. Under random regression imputation, the imputed values y_i^* are given by (8). An imputed estimator of $F_N(t)$ is then given by

$$\hat{F}_I(t) = \sum_{i \in S} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i \in S} \tilde{d}_i (1 - r_i) 1(y_i^* \leq t). \quad (29)$$

Under balanced random regression imputation, the imputed values y_i^{**} are given by (16). Under this imputation scheme, an imputed estimator of $F_N(t)$ is given by

$$\hat{F}_{BALI}(t) = \sum_{i \in S} \tilde{d}_i r_i 1(y_i \leq t) + \sum_{i \in S} \tilde{d}_i (1 - r_i) 1(y_i^{**} \leq t). \quad (30)$$

We assume that the following regularity conditions hold :

C1 : $\max \tilde{\omega}_i = O(1/n)$;

C2 : $\max \tilde{d}_i = O(1/n)$;

C3 : There exists a constant κ such that $\kappa < p_i \equiv Pr(r_i = 1)$;

C4 : $(\hat{B}_r, \hat{\sigma}) \rightarrow_{\mathbb{P}} (\beta, \sigma)$, where $\rightarrow_{\mathbb{P}}$ stands for the convergence in probability;

C5 : $F_{\epsilon}(\cdot)$ is absolutely continuous;

C6 : The components of the vector of auxiliary variables z_i as well as the number K of auxiliary variables are bounded.

The assumptions (C1) and (C2) guarantee that no extreme weight dominates the others. The assumption (C3) states that the response probability is bounded away from 0. The assumption (C4) states that both \hat{B}_r and $\hat{\sigma}$ are consistent estimators of β and σ , respectively.

Theorem 1 *Suppose that conditions C1-C6 hold. If the imputation model (3) holds, then $\hat{F}_I(t) \rightarrow_{\mathbb{P}} F_N(t)$.*

Proof : The proof is given in the Appendix.

Theorem 2 *Suppose that conditions C1-C6 hold. If the imputation model (3) holds, then $\hat{F}_{BALI}(t) \rightarrow_{\mathbb{P}} F_N(t)$.*

Proof : The proof is given in the Appendix.

6 The case of a categorical variable

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 modelled; that is,

$$\phi_i^k \equiv pr(y_i = k) = p(z_i; \gamma^{0,k})$$

where $\sum_{k=1}^K \phi_i^k = 1$, for some function $p(z_i; \cdot)$ with parameter γ evaluated at $\gamma^{0,k}$ where 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

$$\hat{\phi}_i^k = p(z_i; \hat{\gamma}^k) \quad (31)$$

be the estimated probability for unit i of possessing the characteristic of interest k , where $\sum_{k=1}^K \hat{\phi}_i^k = 1$. For simplicity, we consider only the estimation of the proportion of individuals that possess a characteristic k , $\bar{Y}_k = N^{-1} \sum_{i \in U} 1(y_i = k)$. An imputed estimator of \bar{Y}_k is given by $\bar{y}_{kI} = N^{-1} \hat{Y}_{kI}$, where \hat{Y}_{kI} is given by (2) where $x_i = 1$ the variable y_i is replaced by $1(y_i = k)$. For missing y_i , we use

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

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

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

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

$$\sum_{i \in s} d_i (1 - r_i) 1(y_i^{**} = k) = \sum_{i \in s} d_i (1 - r_i) \hat{\phi}_i^k \text{ for any } k = 1, \dots, K. \quad (32)$$

An adaptation of the Cube algorithm may be used so that equations (32) are approximately satisfied; see Chauvet, Deville and Haziza (2010) for more details.

7 Simulation Study

We conducted a limited simulation study to test the performance of the procedures described in sections 2 and 3. We first generated 2 finite populations

of size $N = 10,000$, each containing one variable of interest, y and one auxiliary variable z . First, in each population, the variable z was generated from a Gamma distribution with shape and scale parameters equal to 2 and 5, respectively. Then, the y -values were generated given the z -values according to the model

$$y_i = \beta z_i + \sqrt{z_i} \eta_i. \quad (33)$$

The parameter β was set to 1, whereas the η_i 's were generated according to a normal distribution with mean 0 and variance σ^2 . The parameter σ^2 was chosen to lead to a coefficient of determination (R^2) approximately equal to 0.36 (for population 1) and 0.64 (for population 2).

Our objective is to estimate two parameters : (i) the population mean of the y -values, $\bar{Y} = N^{-1} \sum_{i \in U} y_i$ and (ii) the finite population distribution function, $F_N(t) = N^{-1} \sum_{i \in U} 1(y_i \leq t)$ for different values of t (0.25 ; 0.5 ; 0.75). From each population, we selected 1,000 samples of size $n = 500$ by means of rejective sampling also called Conditional Poisson Sampling (e.g., Hajek, 1964 and Tillé, 2006) with inclusion probabilities, π_i , proportional to z . That is, we have $\pi_i = nz_i/Z$, where $Z = \sum_{i \in U} z_i$. From each generated sample, we generated nonresponse to the variable y according to a uniform response mechanism. Also, units respond independently from one another. The response indicators r_i for $i \in s$ were then generated independently 1,000 times from a Bernoulli distribution with parameter p_0 equal to 0.5 (respectively, 0.75), which led to 1,000 sets of respondents. In each sample containing respondents and nonrespondents, imputation was performed according to three methods : (i) deterministic ratio imputation, (ii) random ratio imputation and (iii) random balanced ratio imputation. All three methods are motivated by the imputation model (3) with z_i scalar and $v_i = z_i$. For deterministic ratio imputation, the imputed values are given by (5) with $\epsilon_i^* = 0$ for all i . The imputed values for random ratio imputation and random balanced ratio imputation are respectively given by (5) and (14).

Then, we computed : (i) the imputed estimator of \bar{Y} given by $\hat{\theta}_I$ in (2) with $x_i = 1$ for all i and (ii) the imputed estimator of $F(t)$ given by $\hat{F}_I(t)$ in (29). As a measure of the bias of $\hat{\theta}_I$, we used the monte carlo percent relative bias (RB) given by

$$\text{RB}(\hat{\theta}_I) = \frac{E_{MC}(\hat{\theta}_I) - \theta_N}{\theta_N} \times 100, \quad (34)$$

where $E_{MC}(\hat{\theta}_I) = \sum_{r=1}^{1,000} \hat{\theta}_I^{(r)} / 1000$, and $\hat{\theta}_I^{(r)}$ denotes the estimator $\hat{\theta}_I$ in the r -th sample, $r = 1, \dots, 1,000$. As a measure of variability of $\hat{\theta}_I$, we used the Monte Carlo Mean Square Error (MSE) given by

$$\text{MSE}(\hat{\theta}_I) = \frac{1}{1,000} \sum_{r=1}^{1,000} (\hat{\theta}_I^{(r)} - \theta_N)^2. \quad (35)$$

Let $\hat{\theta}_I^{(RI)}$, $\hat{\theta}_I^{(RRI)}$, and $\hat{\theta}_I^{(RBRI)}$ denote the estimator $\hat{\theta}_I$ under ratio imputation, random ratio imputation and random balanced ratio imputation, respectively. In order to compare the relative stability of the imputed estimators, using $\hat{\theta}_I^{(RRI)}$ as the reference, we used the following measure :

$$\text{RE} = \frac{\text{MSE}(\hat{\theta}_I^{(\cdot)})}{\text{MSE}(\hat{\theta}_I^{(RRI)})}. \quad (36)$$

Monte carlo measures for $\hat{F}_I(t)$ are obtained from (34)-(36) by replacing $\hat{\theta}_I$ with $\hat{F}_I(t)$ and θ_N with $F_N(t)$.

Table 1 shows the values of RB and RE corresponding to the imputed estimator $\hat{\theta}_I$. It is clear from Table 1 that $\hat{\theta}_I$ is approximately unbiased in all the scenarios, as expected. In terms of RE, results show that $\hat{\theta}_I^{(RI)}$ has the smallest MSE. This result is not surprising since the imputation variance is identically equal to zero for deterministic imputation methods. Also, we note that $\hat{\theta}_I^{(RBRI)}$ is more efficient than $\hat{\theta}_I^{(RRI)}$ with values of RE ranging from 0.82 to 0.93. Finally, we note that $\hat{\theta}_I^{(RBRI)}$ is slightly less efficient than $\hat{\theta}_I^{(RI)}$. This is due to the fact that the balancing equations needed to be relaxed in the last phase of the imputation

		Deterministic ratio imputation	Random ratio imputation	Random balanced ratio imputation
$p_0 = 0.5$				
Population 1	RB	-0.01	0.43	0.10
	RE	0.78	1	0.82
Population 2	RB	0.51	0.72	0.73
	RE	0.81	1	0.85
$p_0 = 0.75$				
Population 1	RB	0.40	0.54	0.52
	RE	0.85	1	0.87
Population 2	RB	0.73	0.80	0.87
	RE	0.90	1	0.93

TABLE 1 – Monte carlo percent relative bias of the imputed estimator and relative efficiency

algorithm in order to end the selection of residuals for non-responding units. As a result, equation (13) did not hold exactly and so the imputation variance was not entirely eliminated.

We now turn to the distribution function, $F_N(t)$. Table 2 shows the RB of the imputed estimator and the RE, corresponding to the imputed estimator $\hat{F}_I(t)$. As expected, the estimators of quantiles under ratio imputation are considerably biased, with absolute RB ranging from 1.49 to 50.13. In terms of relative bias, both random ratio imputation and random balanced ratio imputation show almost not bias, except when $F(t) = 0.25$ for random balanced ratio imputation. These results can be explained by the fact that both imputation methods succeed in preserving the distribution of the variable y . Also, we note that the imputed estimator $\hat{F}_I(t)$ under random balanced ratio imputation is slightly more efficient than the corresponding estimator under random ratio imputation in all the scenarios with a value of RE varying from 0.89 to 1.00. Our results are consistent with those obtained by Chen, Rao and Sitter (2000).

			Deterministic ratio imputation	Random ratio imputation	Random balanced ratio imputation
			$p_0 = 0.5$		
Population 1	0.25	RB	-50.13	0.00	-2.71
		RE	18.41	1	1.00
	0.50	RB	-3.41	0.21	0.09
		RE	2.51	1	0.96
	0.75	RB	10.80	-0.30	0.25
		RE	11.38	1	0.90
Population 2	0.25	RB	-44.57	-0.46	-2.47
		RE	16.92	1	0.92
	0.50	RB	-2.87	0.23	0.04
		RE	1.51	1	0.89
	0.75	RB	6.11	-0.11	0.24
		RE	4.61	1	0.89
			$p_0 = 0.75$		
Population 1	0.25	RB	-24.95	0.13	-0.95
		RE	8.25	1	0.98
	0.50	RB	-1.49	0.10	0.11
		RE	1.51	1	0.98
	0.75	RB	5.53	0.03	0.14
		RE	4.56	1	0.92
Population 2	0.25	RB	-22.27	-0.43	-1.33
		RE	6.79	1	1.00
	0.50	RB	-1.82	-0.21	-0.26
		RE	1.26	1	0.95
	0.75	RB	2.92	-0.13	-0.01
		RE	2.11	1	0.92

TABLE 2 – Monte Carlo percent relative bias of the imputed estimator of the distribution function and relative efficiency

8 Summary and Discussion

In this paper, we have studied the use of balanced random imputation methods 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 described in 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. In the case of regression imputation, We showed that the proposed method preserves the distribution of the variables being imputed. Results from a limited simulation study showed that the proposed balanced random imputation method was almost as efficient as the corresponding deterministic imputation method in all the scenarios.

In this paper, we have not considered the problem of variance estimation in the context of balanced random imputation. In the context of deterministic and random regression imputation, several variance estimation methods are available; e.g., Rao and Shao (1992), Särndal (1992) and Shao and Steel (1999). For balanced random imputation, the problem of variance estimation becomes more complex since the Cube algorithm used for selecting the random residuals involves a rounding process called the landing phase. Correct variance estimation requires to estimate 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.

Acknowledgement

David Haziza's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors wish to thank two referees for useful comments and suggestions that helped improving the quality of the paper significantly.

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 Joint Statistical 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.

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.

- Haziza, D. and Rao, J.N.K.(2005), Inference for domains under imputation for missing data. *The Canadian Journal of Statistics*, 33, pp. 149-161.
- 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.
- 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.
- Praskova, Z. (1995), On Hajek's conjecture in stratified sampling. *Kybernetika*, 31, pp. 303- 314.
- 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.
- Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York.

A Proof of Theorem 1

First note that the total error of $\hat{F}_I(t)$ may be written as

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

Under standard regularity conditions (e.g., Isaki et Fuller, 1981), we have $\hat{F}_N(t) - F_N(t) = O_p(n^{-1/2})$. Consequently, it suffices to show that

$$\hat{F}_I(t) - \hat{F}_N(t) \rightarrow_{\mathbb{P}} 0. \quad (37)$$

To prove the consistency of the estimated distribution function, it is helpful to separate the different sources of error involved in that estimation. More precisely, the imputed values y_i^* may be obtained in the following way :

1. For all units $i \in s_m$, select independent random residuals $\hat{\epsilon}_i$ from the set $G_r = \{\epsilon_j; j \in s_r\}$ of exact residuals. Let $j(i)$ denote the selected donor for unit i , and $\hat{y}_i = z_i^\top \beta + \sigma \sqrt{v_i} \hat{\epsilon}_i$.
2. In \hat{y}_i , the residual $\hat{\epsilon}_i$ is typically unknown. We replace it with the estimated residual $\tilde{\epsilon}_i = \tilde{\epsilon}_{j(i)} = \hat{\sigma}^{-1} v_{j(i)}^{-1/2} (y_{j(i)} - z_{j(i)}^\top \hat{B}_r)$. Let $\tilde{y}_i = z_i^\top \beta + \sigma \sqrt{v_i} \tilde{\epsilon}_i$.
3. In \tilde{y}_i , both β and σ need to be estimated to get the final, imputed value $y_i^* = z_i^\top \hat{B}_r + \hat{\sigma} \sqrt{v_i} \tilde{\epsilon}_i$.

Consequently, we may write

$$\hat{F}_I(t) - \hat{F}_N(t) = T_1 + T_2 + T_3, \quad (38)$$

where

$$T_1 = \sum_{i \in s} \tilde{d}_i (1 - r_i) \{1(\hat{y}_i \leq t) - 1(y_i \leq t)\},$$

$$T_2 = \sum_{i \in s} \tilde{d}_i (1 - r_i) \{1(\tilde{y}_i \leq t) - 1(\hat{y}_i \leq t)\},$$

and

$$T_3 = \sum_{i \in s} \tilde{d}_i (1 - r_i) \{1(y_i^* \leq t) - 1(\tilde{y}_i \leq t)\}.$$

The term T_1 represents the error due to the random selection of a donor for imputation of missing y_i . The term T_2 represents the error due to the estimation of the unknown residual of the selected donor. Finally, the term T_3 represents the error due to the estimation of the unknown parameters of the imputation model (3).

We first show that $T_1 \rightarrow_{\mathbb{P}} 0$. First note that, for any $\delta > 0$, the Bienaymé-Chebyshev inequality implies that

$$pr(|T_1| > \delta) \leq \delta^{-2} V_{mI}(T_1), \quad (39)$$

where $V_{mI}(\cdot)$ denotes the variance under both the imputation model and the imputation mechanism. Also, we have

$$T_1 = \sum_{i \in s} \tilde{d}_i (1 - r_i) \{1(\hat{\epsilon}_i \leq t_i) - 1(\epsilon_i \leq t_i)\}$$

where $t_i = \sigma^{-1} v_i^{-1/2} (t - z_i^\top \beta)$, and

$$V_{mI}(T_1) = E_m V_I(T_1) + V_m E_I(T_1). \quad (40)$$

We focus in the first term in the right-hand side of (40) first. We have

$$\begin{aligned} V_I(T_1) &= V_I \left[\sum_{i \in s} \tilde{d}_i (1 - r_i) 1(\hat{\epsilon}_i \leq t_i) \right] \\ &= \sum_{i \in s} \tilde{d}_i^2 (1 - r_i) V_I [1(\hat{\epsilon}_i \leq t_i)] \\ &= \sum_{i \in s} \tilde{d}_i^2 (1 - r_i) \hat{F}_\epsilon(t_i) (1 - \hat{F}_\epsilon(t_i)) \end{aligned} \quad (41)$$

where $\hat{F}_\epsilon(t) = \sum_{j \in s} \tilde{\omega}_j r_j 1(\epsilon_j \leq t)$. The second line in (41) is a consequence of the independent drawings of the residuals $\hat{\epsilon}_i$. Then, $E_m V_I(T_1)$ is $O(n^{-1})$ from assumption C2. We now turn to the second term in the right-hand side of (40).

We have

$$\begin{aligned}
E_I(T_1) &= \sum_{i \in s} \tilde{d}_i(1-r_i) \left\{ \hat{F}_\epsilon(t_i) - 1(\epsilon_i \leq t_i) \right\} \\
&= \sum_{j \in s} \tilde{\omega}_j r_j \left[\sum_{i \in s} \tilde{d}_i(1-r_i) \right] 1(\epsilon_j \leq t_i) \\
&\quad - \sum_{i \in s} \tilde{d}_i(1-r_i) 1(\epsilon_i \leq t_i),
\end{aligned} \tag{42}$$

and

$$\begin{aligned}
V_m E_I(T_1) &= \sum_{j \in s} \tilde{\omega}_j^2 r_j \left[\sum_{i \in s} \tilde{d}_i(1-r_i) \right]^2 \hat{F}_\epsilon(t_i)(1-\hat{F}_\epsilon(t_i)) \\
&\quad + \sum_{i \in s} \tilde{d}_i^2(1-r_i) \hat{F}_\epsilon(t_i)(1-\hat{F}_\epsilon(t_i))
\end{aligned} \tag{43}$$

since the two terms in the right-hand side of (42) are independent. From assumptions *C1* and *C2*, $V_m E_I(T_1)$ is $O(n^{-1})$. Consequently, $V_m I(T_1)$ is $O(n^{-1})$ and from (39), $T_1 \rightarrow_{\mathbb{P}} 0$.

Then, we establish that $T_2 \rightarrow_{\mathbb{P}} 0$. Let us fix some $\delta > 0$, and let $\eta > 0$ to be specified later. We have :

$$\begin{aligned}
E_I(|T_2|) &\leq E_I \left\{ \sum_{i \in s} \tilde{d}_i(1-r_i) |1(\tilde{y}_i \leq t) - 1(\hat{y}_i \leq t)| \right\} \\
&= \sum_{i \in s} \tilde{d}_i(1-r_i) \sum_{j \in s} \tilde{\omega}_j r_j |1(\tilde{\epsilon}_j \leq t_i) - 1(\epsilon_j \leq t_i)| \\
&= T'_2.
\end{aligned}$$

We note $A_\eta = \{ \|(\hat{B}_r, \hat{\sigma}) - (\beta, \sigma)\| \geq \eta \}$ and $B_\eta = \{ \|(\hat{B}_r, \hat{\sigma}) - (\beta, \sigma)\| < \eta \}$.

Then

$$T'_2 1(A_\eta) \leq \left[\sum_{i \in s} \tilde{d}_i(1-r_i) \right] 1(A_\eta),$$

where the term $[\sum_{i \in s} \tilde{d}_i(1-r_i)]$ is bounded, and $E_I\{1(A_\eta)\} \rightarrow 0$ from assumption *C4*. Consequently,

$$E_I\{T'_2 1(A_\eta)\} \rightarrow 0. \tag{44}$$

On the other hand, we have $1(\tilde{e}_j \leq t_i) = 1(\epsilon_j \leq t_i + \delta_{ij})$, with

$$\delta_{ij} = \sigma^{-1}(\hat{\sigma} - \sigma)t_i - \sigma^{-1}v_i^{-1/2}\{z_j^\top(\beta - \hat{B}_r)\}.$$

Under the event B_η , $|\hat{\sigma} - \sigma| \leq \eta$ and $|\beta - \hat{B}_r| \leq \eta$. Moreover, assumption C6 implies that there exists some constant M_1 such that for any (i, j) we have $|\delta_{ij}| \leq M_1\eta$. Therefore, we get

$$T_2' 1(B_\eta) \leq \sum_{i \in s} \tilde{d}_i(1 - r_i) \sum_{j \in s} \tilde{d}_j r_j 1(t_i - M_1\eta \leq \epsilon_j \leq t_i + M_1\eta),$$

and

$$E_I\{T_2' 1(B_\eta)\} \leq \sum_{i \in s} \tilde{d}_i(1 - r_i) \sum_{j \in s} \tilde{d}_j r_j (F_\epsilon(t_i + M_1\eta) - F_\epsilon(t_i - M_1\eta)).$$

From assumption C5, we may choose η such that $|t - u| \leq 2M_1\eta$ implies that $|F(t) - F(u)| \leq \delta$. Then we obtain

$$E_I\{T_2' 1(B_\eta)\} \leq \delta \sum_{i \in s} \tilde{d}_i(1 - r_i).$$

Since the term $\sum_{i \in s} \tilde{d}_i(1 - r_i)$ is bounded, and since δ may be chosen arbitrarily small, we obtain

$$E_I\{T_2' 1(B_\eta)\} \rightarrow 0. \quad (45)$$

Finally, equations (44) and (45) imply that $E_m E_I |T_2| \rightarrow 0$, and from the Markov inequality $T_2 \rightarrow_{\mathbb{P}} 0$.

It remains to prove that $T_3 \rightarrow_{\mathbb{P}} 0$. We have :

$$\begin{aligned} E_I(|T_3|) &\leq E_I \left\{ \sum_{i \in s} \tilde{d}_i(1 - r_i) |1(y_i^* \leq t) - 1(\tilde{y}_i \leq t)| \right\} \\ &= \sum_{i \in s} \tilde{d}_i(1 - r_i) \sum_{j \in s} \tilde{\omega}_j r_j |1(z_i^\top \hat{B}_r + \hat{\sigma} \sqrt{v_i} \tilde{e}_j \leq t) - 1(z_i^\top \beta + \sigma \sqrt{v_i} \tilde{e}_j \leq t)| \\ &= \sum_{i \in s} \tilde{d}_i(1 - r_i) \sum_{j \in s} \tilde{\omega}_j r_j |1(\epsilon_j \leq t_{1,ij}) - 1(\epsilon_j \leq t_{2,ij})| \end{aligned}$$

where

$$t_{1,ij} = \sigma^{-1}v_i^{-1/2}(t - z_i^\top \hat{B}_r) + \sigma^{-1}v_j^{-1/2}\{z_j^\top (\hat{B}_r - \beta)\}$$

and

$$t_{2,ij} = (\sigma^{-1}\hat{\sigma})\sigma^{-1}v_i^{-1/2}(t - z_i^\top \beta) + \sigma^{-1}v_j^{-1/2}\{z_j^\top (\hat{B}_r - \beta)\}.$$

We have

$$t_{2,ij} - t_i = \sigma^{-1}(\hat{\sigma} - \sigma)t_i + \sigma^{-1}v_j^{-1/2}\{z_j^\top (\hat{B}_r - \beta)\}$$

and

$$t_{1,ij} - t_i = \sigma^{-1}v_i^{-1/2}\{z_i^\top (\beta - \hat{B}_r)\} + \sigma^{-1}v_j^{-1/2}\{z_j^\top (\hat{B}_r - \beta)\}.$$

Like for the term T_2 , the assumptions imply that we may find some constant M_2 such that, under the event B_η , we have for any i and j

$$|t_{1,ij} - t_i| \leq M_2 \eta \text{ and } |t_{2,ij} - t_i| \leq M_2 \eta.$$

A similar proof implies that $T_3 \rightarrow_{\mathbb{P}} 0$.

B Proof of Theorem (2)

We follow the same lines as in the previous proof. So as to prove that

$$\hat{F}_{BALI}(t) - \hat{F}_N(t) \rightarrow_{\mathbb{P}} 0,$$

it is helpful to separate the different sources of error. More precisely, the imputed values y_i^{**} may be obtained in the following way :

1. For all units $i \in s_m$, select random residuals $\hat{\epsilon}_i$ from the set $G_r = \{\epsilon_j; j \in s_r\}$ of exact residuals. The residuals are selected by means of balanced random imputation, that is, in order to satisfy the equation (13). Let $j(i)$ denote the selected donor for unit i , and $\hat{y}_i = z_i^\top \beta + \sigma\sqrt{v_i}\hat{\epsilon}_i$.
2. In \hat{y}_i , the residual $\hat{\epsilon}_i$ is typically unknown. We replace it with the estimated

residual $\tilde{\epsilon}_i = \tilde{e}_{j(i)} = \hat{\sigma}^{-1} v_{j(i)}^{-1/2} (y_{j(i)} - z_{j(i)}^\top \hat{B}_r)$. Let $\tilde{y}_i = z_i^\top \beta + \sigma \sqrt{v_i} \tilde{\epsilon}_i$.

3. In \tilde{y}_i , both β and σ need to be estimated to get the final, imputed value

$$y_i^* = z_i^\top \hat{B}_r + \hat{\sigma} \sqrt{v_i} \epsilon_i^*.$$

Consequently, we may write

$$\hat{F}_{BALI}(t) - \hat{F}_N(t) = T_1 + T_2 + T_3, \quad (46)$$

where the terms in the right-hand side of (46) are the same as in (38), and may be interpreted similarly. Since the steps 2 and 3 involved in the computation of the imputed values are the same in case of the two random imputation mechanisms, the same arguments as previously lead to $T_2 \xrightarrow{\mathbb{P}} 0$ and $T_3 \xrightarrow{\mathbb{P}} 0$. Also, if $V_{m,BALI}(\cdot)$ denotes the variance under both the imputation model and the balanced imputation mechanism, we have

$$V_{m,BALI}(T_1) = E_m V_{BALI}(T_1) + V_m E_{BALI}(T_1) \quad (47)$$

and since $E_{BALI}(T_1) = E_I(T_1)$, the second term in the right-hand side of (47) is $O(n^{-1})$. Consequently, we only need to prove that $E_m V_{BALI}(T_1)$ is $O(n^{-1})$. We note that the random regression imputation mechanism (8) may be seen as a particular case of the balanced random regression imputation mechanism (16), when the balancing constraint (13) is not used. That is, both the random regression imputation and the balanced random regression imputation are performed with the constraints that exactly one donor per non-respondent is selected, but in case of the balanced imputation, the constraint (13) is added. Under exact balancing, the variance formula (8) in Deville and Tillé, p. 574 (see also Praskova, 1995) implies, roughly speaking, that adding extra balancing constraints makes the residuals smaller and consequently decreases the variance. Therefore, up to negligible terms,

$$V_{BALI}(T_1) \leq V_I(T_1),$$

so that the result follows by application of (41).