Preserving the distribution function in surveys in case of imputation for zero inflated data

Guillaume Chauvet, Brigitte Gelein

Preserving the distribution function in surveys in case of imputation for zero inflated data

Brigitte Gelein*and Guillaume Chauvet†

September 22, 2018

Abstract

Item non-response in surveys is usually handled by single imputation, whose main objective is to reduce the non-response bias. Imputation methods need to be adapted to the study variable. For instance, in business surveys, the interest variables often contain a large number of zeros. Motivated by a mixture regression model, we propose two imputation procedures for such data and study their statistical properties. We show that these procedures preserve the distribution function if the imputation model is well specified. The results of a simulation study illustrate the good performance of the proposed methods in terms of bias and mean square error.

Keywords: balanced imputation, imputation model, item non response, mixture model, regression imputation.

*ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France
†ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France

1 Introduction

Item non-response may affect the quality of the estimates when the respondents and the non-respondents exhibit different characteristics with respect to the variables of interest. Item non-response in surveys is usually handled by single imputation, whose main objective is to reduce the non-response bias. The imputation model approach (IM) is commonly used to treat item non-response. It consists in modeling the relationship between the variable of interest and the available auxiliary variables. Single imputation consists of replacing a missing value with an artificial one, obtained by mimicking the imputation model. It leads to a single imputed data set, constructed so that it is possible to apply complete data estimation procedures for obtaining point estimates. The response indicators are therefore not required.

Imputation methods need to be adapted to the study variable. For instance, in business surveys, the interest variables often contain a large number of zeros. In the Capital Expenditure Survey conducted at Statistics Canada, approximately 70% of businesses reported a value of zero to Capital Machinery and 50% reported a value of zero to Capital Construction (Haziza et al., 2014). In case of some interest variable containing a large amount of zeroes, Haziza et al. (2014) propose imputation methods based on a mixture regression model. They prove that these methods lead to doubly robust estimators of the population mean, i.e. the imputed estimator of the mean is consistent whether the interest variable or the non-response mechanism is adequately modeled. However, these methods are not appropriate when estimating more

complex parameters such as the population distribution function.

In this work, we consider estimating the population distribution function in case of imputation for zero inflated data. We use the IM approach, without explicit assumptions on the non-response mechanism for the interest variable. We propose a random imputation method which leads to a consistent estimator of the total and of the distribution function. As recalled in Haziza et al. (2014), random imputation methods usually suffer from an additional variability due to the imputation variance. Therefore, we also propose a balanced version of our method, which enables to reduce the imputation variance. Roughly speaking, it consists of randomly generating the imputed values while satisfying appropriate balancing constraints, by using an adaptation of the Cube algorithm (Deville and Tillé, 2004; Chauvet et al., 2011).

The paper is organized as follows. In Section 2, we describe the theoretical set-up and the notation used in the paper. In Section 3, we briefly recall the two imputation procedures proposed by Haziza et al. (2014), and introduce our two proposed imputation methods. In Section 4, we prove that the proposed random imputation procedure yields a consistent estimator of the total and of the distribution function. Variance estimation for the imputed estimator of the total is discussed in Section 5. The results of a simulation study comparing the four procedures and evaluating the proposed variance estimator are presented in Section 6. We conclude in Section 7. All the proofs are given in the Appendix. Some additional simulation results are available in the Supplementary Material.

2 Theoretical set-up

We are interested in some finite population U of size N, with some variable of interest y taking the value y_i for unit $i \in U$. We note $y_U = (y_1, \ldots, y_N)^\top$ for the vector of values for the variable y. We are interested in estimating the total $t_y = \sum_{i \in U} y_i$, and the finite population distribution function

$$F_N(t) = \frac{1}{N} \sum_{i \in U} 1(y_i \leq t) \tag{2.1}$$

where $1(\cdot)$ is the indicator function.

A sample s of size n is selected according to a sampling design $p(.)$, with π_i the first-order inclusion probability in the sample for unit i. We suppose that $\pi_i > 0$ for any unit $i \in U$, and we note $d_i = \pi_i^{-1}$ the design weight. We note $\delta_U = (\delta_1, \ldots, \delta_N)^\top$ for the vector of sample membership indicators. In case of full response, a complete data estimator of t_y is the expansion estimator or Horvitz-Thompson (1952) estimator

$$\hat{t}_{y\pi} = \sum_{i \in s} d_i y_i. \tag{2.2}$$

This estimator is design-unbiased for t_y, in the sense that $E_p(\hat{t}_{y\pi}) = t_y$ with E_p the expectation under the sampling design $p(.)$, conditionally on y_U. We note V_p the expectation under the sampling design $p(.)$. Concerning the population distribution function F_N, plugging into (2.1) the expansion estimators

of the involved totals yields the plug-in estimator

$$\hat{F}_N(t) = \frac{1}{\hat{N}_\pi} \sum_{i \in s} d_i 1(y_i \leq t) \quad \text{with} \quad \hat{N}_\pi = \sum_{i \in s} d_i. \tag{2.3}$$

Under some mild assumptions on the variable of interest and the sampling design (see Deville, 1999; Cardot et al., 2010), $\hat{F}_N(t)$ is approximately unbiased and mean-square consistent for $F_N(t)$.

We now turn to the case when the variable of interest y is subject to missingness. Let r_i be the response indicator, such that $r_i = 1$ if unit i responded to item y, and $r_i = 0$ otherwise. Let p_i be the response probability of some unit i. We note $r_U = (r_1, \ldots, r_N)^\top$ for the vector of response indicators. We assume that each unit responds independently of one another. Let E_q and V_q denote the expectation and variance under the non-response mechanism, conditionally on the vector y_U of population values and on the vector δ_U of sample membership indicators. An imputation mechanism is used to replace some missing value y_i by an artificial value y_i^*. An imputed estimator for t_y based on observed and imputed values is

$$\hat{t}_{yI} = \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i (1 - r_i) y_i^*. \tag{2.4}$$

Similarly, an imputed estimator of the distribution function based on observed and imputed values is

$$\hat{F}_I(t) = \frac{1}{\hat{N}_\pi} \left\{ \sum_{i \in s} d_i r_i 1(y_i \leq t) + \sum_{i \in s} d_i (1 - r_i) 1(y_i^* \leq t) \right\}. \tag{2.5}$$

In comparison with the estimators obtained in (2.2) and (2.3) with complete data, there are two additional random mechanisms involved in the estimators given in (2.4) and (2.5). First, the non-response mechanism leads to observe the values of y for a part of s only. Then, the imputation mechanism is used to replace missing y_i's with artificial values.

The imputation mechanism is motivated by an imputation model, which is a set of assumptions on the variable y subject to missingness. In the context of a zero-inflated variable of interest, we consider the mixture regression model introduced in Haziza et al. (2014). Namely, we assume that

$$y_i \;=\; \eta_i \left\{ z_i^\top \beta + \sqrt{v_i}\epsilon_i \right\}, \tag{2.6}$$

where the η_i's are independent Bernoulli random variables equal to 1 with probability ϕ_i, and equal to 0 otherwise; the ϵ_i's are independent and identically distributed random variables of mean 0, variance σ^2, and with a common distribution function F_ϵ; the parameters β and σ are unknown, and v_i is a known constant. The vector of auxiliary variables z_i is assumed to be known on the whole sample including non-respondents. To sum up, according to the imputation model (2.6) the variable y_i follows a regression model with a probability ϕ_i, and is equal to 0 otherwise. Let E_m et V_m denote respectively the expectation and variance under the imputation model. We suppose that the sampling design is non-informative, in the sample that the vector δ_U of sample membership indicators is independent of $\epsilon_U = (\epsilon_1, \ldots, \epsilon_N)^\top$ and $\eta_U = (\epsilon_1, \ldots, \epsilon_N)^\top$, conditionally on a set of design variables.

In practice, the ϕ_i's are unknown and need to be estimated. We assume that they may be parametrically modeled as

$$\phi_i = f(u_i, \gamma) \tag{2.7}$$

where f is a known function, u_i is a vector of variables recorded for all sampled units, and γ is an unknown parameter. An estimator of ϕ_i is

$$\hat{\phi}_i = f(u_i, \hat{\gamma}_r) \tag{2.8}$$

with $\hat{\gamma}_r$ an estimator of γ computed on the responding units. We assume that η_i and ϵ_i are independent, conditionally on the vectors z_i and u_i.

In this paper, we use the Imputation Model (IM) approach where the inference is made with respect to the imputation model, the sampling design, the response mechanism and the imputation mechanism. This does not require an explicit modeling of the non-response mechanism unlike the Non-response Model approach (Haziza, 1999), but we assume that the data are missing at random, which means that model (2.6) holds for both the respondents and the non-respondents. We note E_I and V_I the expectation and variance under the imputation mechanism, conditionally on the vectors y_U, δ_U and r_U.

3 Imputation methods

In this Section, we first briefly recall in Sections 3.1 and 3.2 the random imputation methods proposed by Haziza et al. (2014) for zero-inflated data. We then introduce the new methods that we propose in Sections 3.3 and 3.4.

3.1 Haziza-Nambeu-Chauvet random imputation

A first proposal of Haziza et al. (2014) is to use the imputation mechanism

$$y_i^* \quad = \quad \eta_i^* \left\{ z_i^\top \hat{B}_r \right\}, \tag{3.1}$$

where the unknown regression parameter β is estimated by

$$\hat{B}_r = \hat{G}_r^{-1} \left(\frac{1}{N} \sum_{i \in s} \omega_i r_i v_i^{-1} z_i y_i \right) \quad \text{with} \quad \hat{G}_r = \frac{1}{N} \sum_{i \in s} \omega_i r_i \hat{\phi}_i v_i^{-1} z_i z_i^\top, \tag{3.2}$$

where ω_i denotes a so called imputation weight, and $\hat{\phi}_i$ is given in (2.8). The η_i^*'s are independently generated, and η_i^* is equal to 1 with the probability $\hat{\phi}_i$, and is equal to 0 otherwise.

There are several possible choices for the imputation weights ω_i. Using a modeling of the response mechanism for the variable y_i, Haziza et al. (2014) propose to choose the imputation weights so that $\hat{t}_{yI}$ is a doubly robust estimator for t_y. This means that the imputed estimator is approximately unbiased for t_y whether the imputation model or the non-response model is adequately specified. Haziza et al. (2014) also prove that the resulting im-

puted estimator is consistent for t_y under either approach.

The random imputation mechanism in (3.1) has three drawbacks. Firstly, it leads to an additional imputation variance due to the η_i^*'s. To overcome this problem, Haziza et al. (2014) proposed a balanced version of their imputation mechanism that is presented in Section 3.2. Secondly, the imputation mechanism in (3.1) does not lead to an approximately unbiased estimator of the distribution function, as will be illustrated in the simulation study conducted in Section 5. Finally, the consistency of the imputed estimator $\hat{t}_{yI}$ relies on an assumption of mean square consistency for $\hat{B}_r$, which may be difficult to prove since the matrix $\hat{G}_r$ can be close to similarity for some samples. Following Cardot et al. (2013) and Chauvet and Do Paco (2018), we introduce in Sections 3.3 and 3.4 a regularized version of $\hat{B}_r$.

3.2 Haziza-Nambeu-Chauvet balanced imputation

The balanced random imputation procedure of Haziza et al. (2014) consists in replacing a missing value with

$$y_i^* \quad = \quad \tilde{\eta}_i^* \left\{ z_i^\top \hat{B}_r \right\}, \tag{3.3}$$

where the $\tilde{\eta}_i^*$'s are not independently generated, but so that the imputation variance of $\hat{t}_{yI}$ is approximately equal to zero. Indeed, the imputation variance of $\hat{t}_{yI}$ is eliminated if the $\tilde{\eta}_i^*$'s are generated so that

$$\sum_{i \in s} d_i (1 - r_i)(\tilde{\eta}_i^* - \hat{\phi}_i)(z_i^\top \hat{B}_r) \quad = \quad 0. \tag{3.4}$$

Haziza et al. (2014) propose a procedure adapted from the Cube method (Deville and Tillé, 2004; Chauvet and Tillé, 2006) which enables to generate the $\tilde{\eta}_i^*$'s so that (3.4) is satisfied, at least approximately. As a result, the imputation variance is eliminated or at least significantly reduced.

This imputation procedure is called balanced random ϕ-regression (BRR_ϕ) imputation by Haziza et al. (2014). They prove that under the BRR_ϕ imputation, an appropriate choice for the imputation weights ω_i leads to a doubly robust estimator for t_y. Also, their empirical results indicate that it performs well in reducing the imputation variance. A drawback of the BRR_ϕ imputation mechanism is that it does not preserve the distribution function of the imputed variable, because it does not take into account the error terms ϵ_i in the imputation model (2.6). This is empirically illustrated in section 5. To overcome this problem, two new imputation procedures are proposed in Sections 3.3 and 3.4.

3.3 Proposed random imputation

The random imputation procedure that we propose consists in mimicking as closely as possible the imputation model (2.6), by replacing some missing y_i with the imputed value

$$y_i^* \quad = \quad \eta_i^* \left\{ z_i^\top \hat{B}_{ar} + \sqrt{v_i}\epsilon_i^* \right\}, \tag{3.5}$$

where $\hat{B}_{ar}$ is a regularized version of $\hat{B}_r$, and η_i^* is a Bernoulli random variable as defined in (3.1). The ϵ_i^*'s are selected independently and with replacement

in the set of observed residuals

$$E_r = \{e_j \ ; \ r_j = 1 \text{ and } \eta_j = 1\} \quad \text{where} \quad e_j = \frac{y_j - z_j^\top \hat{B}_{ar}}{\sqrt{v_j}}, \tag{3.6}$$

with $Pr(\epsilon_i^* = e_j) = \tilde{\omega}_j$ for any $j \in s$ such that $r_j = 1$ and $\eta_j = 1$, where

$$\tilde{\omega}_j \ = \ \frac{\omega_j}{\sum_{k \in s} \omega_j r_k \eta_k}. \tag{3.7}$$

We note

$$\bar{e}_r = \sum_{j \in s} \tilde{\omega}_j r_j \eta_j e_j \quad \text{and} \quad \sigma_{er}^2 = \sum_{j \in s} \tilde{\omega}_j r_j \eta_j (e_j - \bar{e}_r)^2. \tag{3.8}$$

The regularized version of $\hat{B}_r$ is obtained by following the approach in Cardot et al. (2013) and Chauvet and Do Paco (2018). We first write

$$\hat{G}_r \ = \ \sum_{j=1}^{p} \alpha_{jr} v_{jr} v_{jr}^\top, \tag{3.9}$$

with $\alpha_{jr} \geq \ldots \geq \alpha_{pr}$ the non-negative eigenvalues of $\hat{G}_r$, and where $v_{1r}, \ldots, v_{pr}$ are the associated orthonormal vectors. For some given $a > 0$, the regularized versions of $\hat{G}_r$ and $\hat{B}_r$ are

$$\hat{G}_{ar} = \sum_{j=1}^{p} \max(\alpha_{jr}, a) v_{jr} v_{jr}^\top \quad \text{and} \quad \hat{B}_{ar} = \hat{G}_{ar}^{-1} \left(\frac{1}{N} \sum_{i \in s} \omega_i r_i v_i^{-1} z_i y_i \right) \tag{3.10}$$

The regularization leads to a matrix $\hat{G}_{ar}$ which is always invertible, and such that $\|\hat{G}_{ar}^{-1}\| \leq a^{-1}$ with $\|\cdot\|$ the spectral norm.

We prove in Section 4 that $\hat{B}_{ar}$ is a mean-square consistent estimator of β, and that under the proposed imputation procedure the imputed estimator of the total is mean-square consistent for t_y. Also, we prove that the imputed estimator $\hat{F}_I(t)$ is L_1-consistent for the population distribution function. However, this imputation procedure leads to an additional variability for $\hat{t}_{yI}$ due to the imputation variance. Therefore, a balanced version of this imputation procedure is proposed in Section 3.4.

3.4 Proposed balanced imputation

The balanced procedure consists in replacing a missing value with

$$y_i^* = \tilde{\eta}_i^* \left\{ z_i^\top \hat{B}_{ar} + \sqrt{v_i}\tilde{\epsilon}_i^* \right\}, \tag{3.11}$$

where the $\tilde{\eta}_i^*$'s and the $\tilde{\epsilon}_i^*$'s are not independently generated, but so as to eliminate the imputation variance of $\hat{t}_{yI}$. A sufficient condition for this consists in generating the residuals $\tilde{\eta}_i^*$ and $\tilde{\epsilon}_i^*$ so that

$$\sum_{i\in s} d_i(1-r_i)(\tilde{\eta}_i^* - \hat{\phi}_i)(z_i^\top \hat{B}_r^*) = 0, \tag{3.12}$$

$$\sum_{i\in s} d_i(1-r_i)\tilde{\eta}_i^* \sqrt{v_i}\tilde{\epsilon}_i^* = 0. \tag{3.13}$$

This is done in a two-step procedure: first, the $\tilde{\eta}_i^*$'s are generated by means of Algorithm 1 in Haziza et al. (2014), so that (3.12) is approximately respected; then, the $\tilde{\epsilon}_i^*$'s are generated by using Algorithm 1 described in Chauvet et al. (2011), so that (3.13) is approximately respected.

Since the balancing equations (3.12) and (3.13) are usually only approximately respected, the imputation variance is not completely eliminated, but it may be significantly reduced: see the simulation study in Section 5. Though the balanced imputation procedure is expected to provide estimators with smaller variance, the asymptotic properties of these estimators are difficult to study due to intricate dependencies introduced in the imputation process. Extending the results in Section 4 is a challenging problem for further theoretical research.

4 Properties of the proposed methods

To study the asymptotic properties of the sampling designs and estimators, we use the asymptotic framework of Isaki and Fuller (1982). We suppose that the population U belongs to a nested sequence $\{U_\tau\}$ of finite populations with increasing sizes N_τ, and that the vector of values for the variable of interest $y_{U\tau} = (y_{1\tau}, \ldots, y_{N\tau})^\top$ belongs to a nested sequence $\{y_{U\tau}\}$ with increasing sizes N_τ. For simplicity, the index τ is omitted in what follows and all limits are computed when $\tau \to \infty$.

We consider the following regularity assumptions:

H1: Some constants $C_1, C_2 > 0$ exist, s.t. $C_1 \leq Nn^{-1}\pi_i \leq C_2$ for any $i \in U$.

H2: Some constant C_3 exists, s.t. $\sup_{i \neq j \in U} \left(n \left| 1 - \frac{\pi_{ij}}{\pi_i \pi_j} \right| \right) \leq C_3$.

H3: Some constants $C_4, C_4' > 0$ exist, s.t. $C_4 \leq \min_{i \in U} p_i$ and $C_4' \leq \min_{i \in U} \phi_i$.

H4: Some constants $C_5, C_6 > 0$ exist, s.t. $C_5 \leq N^{-1} n\omega_i \leq C_6$ for any $i \in U$.

H5: Some constants $C_7, C_8, C_9 > 0$ exist, s.t. $C_7 \leq v_i \leq C_8$ and $\|z_i\| \leq C_9$ for any $i \in U$. Also, the matrix

$$G = \frac{1}{N} \sum_{i \in U} \omega_i \pi_i p_i \phi_i v_i^{-1} z_i z_i^\top \tag{4.1}$$

is invertible, and the constant a chosen is s.t. $\|G^{-1}\| \leq a^{-1}$.

H6: We have $E\left(\|\hat{\gamma}_r - \gamma\|^2\right) = O(n^{-1})$.

H7: Some constant C_{11} exists, s.t. for any vector $\tilde{\gamma}$

$$|f(u_i, \tilde{\gamma}) - f(u_i, \gamma)| \leq C_{11} \|\tilde{\gamma} - \gamma\| \text{ for all } i \in U.$$

It is assumed in (H1) that the inclusion probabilities do not differ much from that obtained under simple random sampling, so that no design weight dominates the other. It is assumed in (H2) that the units in the population are not far from being independently selected: this assumption is verified for stratified simple random sampling and rejective sampling (Hájek, 1964), for example. It is assumed in (H3) that the response probabilities are bounded away from 0, i.e. there is no hard-core non-respondents, and that the probabilities of observing a null value are also bounded away from 0, i.e. the variable of interest is not degenerate. The assumption (H4) is related to the imputation weights, and is similar to assumption (H1). The assumption (H5) is related to the imputation model, and is necessary to control the behaviour of the regularized estimator $\hat{B}_{ar}$; see Cardot et al. (2013) and Chauvet and

Do Paco (2018). It is assumed in (H6) that the estimator $\hat{\gamma}_r$ is $\sqrt{n}$ mean-square consistent for the parameter γ. This assumption is somewhat strong, but is needed to obtain the standard rate of convergence for the imputed estimator of the total. It is assumed in (H7) that $f(\cdot, \cdot)$ is Lipschitz-Continuous in its second component. The assumptions (H5) and (H6) are also considered in Haziza et al. (2014).

Proposition 1. *Suppose that the imputation model in (2.6) holds and that the assumptions (H1)-(H7) are satisfied. Then we have*

$$E\left\{\|\hat{B}_{ar} - \beta\|^2\right\} = O(n^{-1}). \tag{4.2}$$

Proposition 2. *Suppose that the imputation model in (2.6) holds and that the assumptions (H1)-(H7) are satisfied. Then under the random imputation mechanism proposed in Section 3.3, we have*

$$E\left[\left\{N^{-1}(\hat{t}_{yI} - t_y)\right\}^2\right] = O(n^{-1}). \tag{4.3}$$

Proposition 3. *Suppose that the imputation model in (2.6) holds and that the assumptions (H1)-(H7) are satisfied. Also, suppose that the distribution function F_ϵ is absolutely continuous. Then under the random imputation mechanism proposed in Section 3.3, we have for any $t \in \mathbb{R}$*

$$E\left[\left\{\hat{F}_I(t) - F_N(t)\right\}^2\right] = o(1). \tag{4.4}$$

5 Variance estimation

We now consider variance estimation for the imputed estimator of the total $\hat{t}_{yI}$, under the proposed imputation procedures. The variance estimators are adapted from a linearized variance estimator proposed by Kim and Rao (2009, Section 2) for deterministic/random regression imputation. They are obtained under a variance decomposition which makes use of the reverse approach (Fay, 1996; Shao and Steel, 1999). For simplicity, we suppose that the ϕ_i's are modeled according to a logistic regression model and that the unknown parameter β is the solution of the weighted estimated equation

$$\sum_{i \in s} \omega_i r_i u_i \left\{ \eta_i - f(u_i, \gamma) \right\} \quad = \quad 0, \tag{5.1}$$

with $\text{logit} f(u_i, \gamma) = u_i^\top \gamma$.

5.1 Balanced imputation procedure

We first consider the balanced imputation procedure proposed in Section 3.4. We do not need to account for the imputation variance, since it is approximately eliminated for the estimation of the total with the proposed imputation procedure. By following the approach of Kim and Rao (2009), we obtain after some algebra the two-term variance estimator

$$\hat{V}_{BMRR}(\hat{t}_{yI}) \quad = \quad \hat{V}_1(\hat{t}_{yI}) + \hat{V}_2(\hat{t}_{yI}), \tag{5.2}$$

see equations (10) and (13) in Kim and Rao (2009). The first term in the right-hand side of (5.2) is

$$\begin{aligned}
\hat{V}_1(\hat{t}_{yI}) &= \sum_{i,j\in s}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\right)\hat{\xi}_i\hat{\xi}_j, \\
\text{with} \quad \hat{\xi}_i &= d_i(\hat{\phi}_i z_i^\top \hat{B}_{ar}) + r_i\left(d_i+\omega_i\hat{\phi}_i v_i^{-1}\hat{a}^\top z_i\right)\left(y_i-\hat{\phi}_i z_i^\top \hat{B}_{ar}\right) \\
&+ r_i\omega_i(\hat{b}-\hat{c})^\top u_i\left(\eta_i-\hat{\phi}_i\right), \qquad (5.3)
\end{aligned}$$

with

$$\begin{aligned}
\hat{a} &= \left(\sum_{i\in s} r_i\omega_i\hat{\phi}_i v_i^{-1} z_i z_i^\top\right)^{-1}\sum_{i\in s} d_i(1-r_i)\hat{\phi}_i z_i, \\
\hat{b} &= \left(\sum_{i\in s} r_i\omega_i\hat{\phi}_i(1-\hat{\phi}_i) u_i u_i^\top\right)^{-1}\sum_{i\in s} d_i(1-r_i)\hat{\phi}_i(1-\hat{\phi}_i)(z_i^\top \hat{B}_{ar})u_i, \qquad (5.4) \\
\hat{c} &= \left(\sum_{i\in s} r_i\omega_i\hat{\phi}_i(1-\hat{\phi}_i) u_i u_i^\top\right)^{-1}\sum_{i\in s} \omega_i r_i v_i^{-1}\hat{\phi}_i(1-\hat{\phi}_i)(z_i^\top \hat{a})(z_i^\top \hat{B}_{ar})u_i,
\end{aligned}$$

and with π_{ij} the probability that units i and j are selected together in the sample. The second term in the right-hand side of (5.2) is

$$\hat{V}_2(\hat{t}_{yI}) = \sum_{i\in s} r_i d_i\left\{(1+\omega_i\pi_i v_i^{-1}\hat{a}^\top z_i)(y_i-\hat{\phi}_i z_i^\top \hat{B}_{ar}) + \omega_i\pi_i(\hat{b}-\hat{c})^\top u_i(\eta_i-\hat{\phi}_i)\right\}^2 \qquad (5.5)$$

As underlined by Kim and Rao (2009), $\hat{V}_2(\hat{t}_{yI})$ is not sensitive to a mis-specification of the covariance structure in model (2.6).

5.2 Random imputation procedure

We now consider the random imputation procedure proposed in Section 3.3. We need to account to the additional variance due to the imputation process. By following once again the approach in Kim and Rao (2009, Section 4.1), we obtain the variance estimator

$$\hat{V}_{MRR}(\hat{t}_{yI}) = \hat{V}_{BMRR}(\hat{t}_{yI}) + \hat{V}_3(\hat{t}_{yI}), \tag{5.6}$$

where $\hat{V}_{BMRR}(\hat{t}_{yI})$ is given in equation (5.2), and with

$$\hat{V}_3(\hat{t}_{yI}) = \sum_{i \in s} d_i^2 (1 - r_i)(y_i^* - \hat{\phi}_i z_i^\top \hat{B}_{ar})^2, \tag{5.7}$$

with y_i^* the imputed value given in equation (3.5).

6 Simulation study

To evaluate the performance of the proposed imputation methods, we implement a simulation study inspired by Haziza et al. (2014). We generate nine finite populations of size $N = 10,000$ with an interest variable y and an auxiliary variable z. The values of z are generated according to a Gamma distribution with shift parameter 2 and scale parameter 5. The values of y are generated according to the following mixture model:

$$y_i = \eta_i(a_0 + a_1 z_i + \epsilon_i), \tag{6.1}$$

where the ϵ_i's are generated according to a standard normal distribution with variance σ^2. We use $a_0 = 30$ and $a_1 = 1.5$. Also, we choose three different values of σ^2 so that the coefficient of determination R^2 equals 0.4, 0.5 or 0.6 for the units i such that $\eta_i = 1$.

The η_i's are generated according to a Bernoulli distribution with parameter ϕ_i, and

$$\log\left(\frac{\phi_i}{1-\phi_i}\right) \quad = \quad b_0 + b_1 z_i, \tag{6.2}$$

and with four possible values for the parameters b_0 and b_1, chosen so that the proportion of non-null values is approximately equal to 0.60, 0.70, or 0.80. The three different proportion of non-null values, crossed with the three different levels for the R^2, lead to the nine finite populations.

In each population, we select $R = 1,000$ samples by means of rejective sampling (Hájek, 1964) of size $n = 500$, with inclusion probabilities proportional to the variable z_i. In each sample, we generate a response indicator r_i for unit i according to a Bernoulli distribution with parameter p_i such that

$$\log\left(\frac{p_i}{1-p_i}\right) \quad = \quad c_0 + c_1 z_i. \tag{6.3}$$

We use three possible values for the parameters c_0 and c_1, chosen so that the proportion of respondents is approximately equal to 0.30, 0.50 or 0.70.

6.1 Properties of point estimators

In this Section, we are interested in estimating the total t_y, and the distribution function $F_N(t)$ with $t = t_\alpha$, the α-th quintile. In this simulation study, we consider the values $\alpha = 0.50, 0.75$ and 0.90. We compare four imputation methods to handle non-response: (i) random imputation (RR_ϕ) proposed by Haziza et al. (2014), and presented in Section 3.1; (ii) balanced random imputation (BRR_ϕ) proposed by Haziza et al. (2014), and presented in Section 3.2; (iii) proposed random imputation method (MRR_ϕ), presented in Section 3.3; (iv) proposed balanced random imputation method ($BMRR_\phi$), presented in Section 3.4. For each of the four methods, we use imputation weights $\omega_i = 1$, and the ϕ_i's and p_i's are estimated by means of logistic regression modeling. In each sample, missing values are replaced by imputed values according to imputation methods (i) to (iv), and the imputed estimators $\hat{t}_{yI}$ and $\hat{F}_I(t_\alpha)$ are computed.

As a measure of bias of an estimator $\hat{\theta}_I$ of a finite population parameter θ, we compute the Monte Carlo percent relative bias

$$RB_{MC}(\hat{\theta}_I) = \frac{100}{R}\sum_{k=1}^{R}\frac{(\hat{\theta}_{I(k)} - \theta)}{\theta}, \tag{6.4}$$

where $\hat{\theta}_{I(k)}$ denotes the imputed estimator computed in the k-th sample. As a measure of relative efficiency for each imputation method, using $BMRR_\phi$

as a benchmark, we computed

$$RE_{MC}(\hat{\theta}_I) = \frac{MSE_{MC}(\hat{\theta}_I)}{MSE_{MC}(\hat{\theta}_{BMRR_\phi})} \quad \text{with} \quad MSE_{MC}(\hat{\theta}_I) = \frac{1}{R}\sum_{k=1}^{R}(\hat{\theta}_{I(k)} - \theta)^2,$$

the Mean Square Error of $\hat{\theta}_I$ approximated by means of the R simulations. We observed no qualitative difference according to the different response rates. For brevity, we therefore only present the simulation results with an average proportion of respondents of 0.50. The simulation results for the two other response rates are given in the Supplementary Material.

We first consider the estimation of the total t_y, for which the simulation results are given in Table 1. The four imputation methods lead to approximately unbiased estimators of the total, as expected. Turning to the relative efficiency (RE), we note that in all studied cases the balanced version of an imputation method outperforms its unbalanced version. Also, the two balanced imputation procedures exhibit similar efficiency, with BRR_ϕ performing slightly better. This is likely due to fact that the balancing equations (3.12) and (3.13) are not exactly respected due to the landing phase of the cube method (see Deville and Tillé, 2004).

We now consider the estimation of the population distribution function, for which the simulation results are presented in Table 2. In all the cases considered, the two proposed imputation methods MRR_ϕ and $BMRR_\phi$ lead to approximately unbiased estimators of the distribution function, with absolute relative biases no greater than 4 % . On the contrary, the RR_ϕ and

		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	1.30	1.06	1.25	0.98	1.20	1.09	1.21	1.00
0.4	0.7	0.21	1.09	0.26	0.95	0.22	1.14	0.23	1.00
0.4	0.8	0.05	1.02	0.09	0.97	0.04	1.11	0.09	1.00
0.5	0.6	1.26	1.04	1.25	0.98	1.22	1.06	1.26	1.00
0.5	0.7	0.25	1.00	0.30	0.98	0.24	1.05	0.34	1.00
0.5	0.8	0.16	1.02	0.06	0.96	0.19	1.09	0.06	1.00
0.6	0.6	1.13	1.10	1.20	0.99	1.15	1.14	1.22	1.00
0.6	0.7	0.25	1.05	0.30	0.96	0.24	1.10	0.27	1.00
0.6	0.8	-0.02	1.04	0.07	0.97	0.00	1.08	0.08	1.00

Table 1: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the total with an average response probability of 50%

the BRR_ϕ imputation methods lead to biased estimators, and the absolute relative bias can be as large as 16 % . We note that the bias is larger for the lower quantiles. Turning to the relative efficiency, we note that MRR_ϕ and $BMRR_\phi$ always outperform RR_ϕ and BRR_ϕ, which is partly due to the bias under these latter imputation methods. Comparing the two proposed imputation methods, we note that $BMRR_\phi$ is equivalent or better than MRR_ϕ in terms of efficiency, with values of RE ranging from 1.00 to 1.12 for MRR_ϕ.

6.2 Properties of variance estimators

We now consider the properties of the variance estimators proposed in Section 5. Under the rejective sampling design used in the simulation study, we replace the component $\hat{V}_1(\hat{t}_{yI})$ given in (5.3) with the Hajek-Rosen variance estimator

$$\hat{V}_{HR,1}(\hat{t}_{yI}) = \frac{n}{n-1}\sum_{i\in s}(1-\pi_i)(\hat{\xi}_i - \hat{R})^2 \text{ with } \hat{R} = \frac{\sum_{i\in s}(1-\pi_i)\hat{\xi}_i}{\sum_{i\in s}(1-\pi_i)}, \quad (6.5)$$

		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
		RB %	RE	RB %	RE	RB %	RE	RB %	RE
R^2	$\overline{\phi}$	50% quartile							
0.4	0.6	-15.64	3.96	-15.56	3.86	-3.75	1.04	-3.84	1.00
0.4	0.7	-18.73	7.85	-18.83	7.74	-1.33	1.09	-1.31	1.00
0.4	0.8	-8.82	4.21	-8.85	4.19	-1.08	1.12	-1.15	1.00
0.5	0.6	-15.00	3.91	-14.99	3.84	-3.54	1.03	-3.66	1.00
0.5	0.7	-15.17	6.30	-15.20	6.34	-0.99	1.03	-0.99	1.00
0.5	0.8	-5.86	2.79	-5.82	2.74	-1.18	1.07	-0.83	1.00
0.6	0.6	-14.46	3.78	-14.48	3.71	-3.75	1.08	-3.89	1.00
0.6	0.7	-12.22	5.03	-12.29	5.04	-0.78	1.06	-0.76	1.00
0.6	0.8	-3.82	2.13	-3.85	2.12	-0.81	1.05	-0.75	1.00
R^2	$\overline{\phi}$	75% quartile							
0.4	0.6	9.81	6.19	9.81	6.20	1.82	1.03	1.80	1.00
0.4	0.7	11.93	6.53	11.94	6.53	3.06	1.02	3.07	1.00
0.4	0.8	10.59	6.83	10.57	6.80	2.25	1.09	2.10	1.00
0.5	0.6	9.14	4.87	9.13	4.86	2.33	1.05	2.35	1.00
0.5	0.7	11.23	4.75	11.25	4.76	3.86	1.03	3.83	1.00
0.5	0.8	9.52	5.28	9.54	5.30	2.58	1.07	2.61	1.00
0.6	0.6	8.55	3.81	8.59	3.81	2.81	1.01	2.89	1.00
0.6	0.7	10.60	3.55	10.57	3.53	4.51	1.01	4.55	1.00
0.6	0.8	8.60	3.85	8.58	3.84	3.09	1.05	2.97	1.00
R^2	$\overline{\phi}$	90% quartile							
0.4	0.6	4.82	2.46	4.80	2.45	2.33	1.00	2.32	1.00
0.4	0.7	5.26	2.46	5.27	2.46	2.75	1.01	2.80	1.00
0.4	0.8	4.87	3.03	4.87	3.03	2.08	1.00	2.06	1.00
0.5	0.6	4.52	1.94	4.54	1.94	2.65	1.02	2.63	1.00
0.5	0.7	5.00	1.93	5.00	1.93	3.14	1.03	3.13	1.00
0.5	0.8	4.53	2.40	4.53	2.40	2.36	1.03	2.34	1.00
0.6	0.6	4.32	1.60	4.32	1.60	2.90	1.01	2.89	1.00
0.6	0.7	4.85	1.58	4.84	1.57	3.45	1.00	3.47	1.00
0.6	0.8	4.23	1.89	4.22	1.89	2.60	1.02	2.58	1.00

Table 2: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 50%, 75% and 90% quartiles with an average response probability of 50%

see also Chauvet and Do Paco (2018). This leads to the simplified variance estimator

$$\tilde{V}_{BMRR}(\hat{t}_{yI}) \quad = \quad \hat{V}_{HR,1}(\hat{t}_{yI}) + \hat{V}_2(\hat{t}_{yI}), \tag{6.6}$$

for the proposed balanced imputation procedure $BMRR_\phi$, and to the simplified variance estimator

$$\tilde{V}_{MRR}(\hat{t}_{yI}) \quad = \quad \tilde{V}_{BMRR}(\hat{t}_{yI}) + \hat{V}_3(\hat{t}_{yI}), \tag{6.7}$$

for the proposed random imputation procedure MRR_ϕ.

We computed the Monte-Carlo percent relative bias of these two variance estimators, using an independent simulation-based approximation of the true mean square error of $\hat{t}_{yI}$ based on $10,000$ simulations. We also computed the coverage rates of the associated normality-based confidence intervals, with nominal error rate of 2.5% in each tail. We only consider the two cases when the average proportion of respondents is 0.50 and 0.70. We first consider the results for $BMRR_\phi$, which are presented in Table 3. The variance estimator $\tilde{V}_{BMRR}(\hat{t}_{yI})$ is approximately unbiased with $\bar{p} = 0.50$, but is slightly negatively biased with $\bar{p} = 0.70$. This is likely due to the fact that the imputation variance is not completely eliminated with the proposed balanced imputation procedure, due to the landing phase of the cube method. The coverage rates are approximately respected in any case, but the confidence intervals tend to be narrow when $\bar{p} = 0.70$ which is in accordance with the variance estimator

	Population 1					
	$\phi = 0.6$		$\phi = 0.7$		$\phi = 0.8$	
	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$
RB (%)	-2.4	-6.5	-0.8	-4.4	-2.1	-4.7
Cov. Rate	94.8	92.7	95.0	93.8	96.0	93.1
	Population 2					
	$\phi = 0.6$		$\phi = 0.7$		$\phi = 0.8$	
	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$
RB (%)	-2.1	-5.8	-0.8	-3.6	-3.9	-3.4
Cov. Rate	94.2	92.7	95.7	93.4	95.4	93.6
	Population 3					
	$\phi = 0.6$		$\phi = 0.7$		$\phi = 0.8$	
	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$
RB (%)	-2.4	-5.5	-2.0	-3.3	-1.7	-3.9
Cov. Rate	94.5	92.4	95.1	93.4	95.1	93.2

Table 3: Monte-Carlo percent relative bias of the variance estimator and coverage rate for the proposed balanced imputation procedure $BMRR_\phi$

being negatively biased. We now turn to MRR_ϕ, for which the simulation results are presented in Table 4. The variance estimator $\tilde{V}_{BMRR}(\hat{t}_{yI})$ is approximately unbiased with $\bar{p} = 0.70$, but is slightly positively biased with $\bar{p} = 0.50$. The coverage rates are approximately respected in all cases.

7 Conclusion

In this paper, we considered imputation for zero-inflated data. We proposed two imputation methods which enable to respect the nature of the data, and in particular which preserve the finite population distribution function. In particular, we proposed a balanced imputation method which enables to preserve the distribution of the imputed variable while being fully efficient for the estimation of a total.

	Population 1					
	$\phi = 0.6$		$\phi = 0.7$		$\phi = 0.8$	
	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$
RB (%)	4.7	1.8	7.6	3.0	5.8	3.3
Cov. Rate	95.2	94.2	96.2	94.6	96.3	94.1
	Population 2					
	$\phi = 0.6$		$\phi = 0.7$		$\phi = 0.8$	
	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$
RB (%)	6.5	3.2	8.0	2.5	4.7	3.6
Cov. Rate	95.4	93.8	96.4	94.0	94.8	93.7
	Population 3					
	$\phi = 0.6$		$\phi = 0.7$		$\phi = 0.8$	
	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$	$\bar{p} = 0.5$	$\bar{p} = 0.7$
RB (%)	6.1	1.4	8.4	3.0	4.1	3.7
Cov. Rate	94.9	94.1	95.7	94.7	96.1	94.4

Table 4: Monte-Carlo percent relative bias of the variance estimator and coverage rate for the proposed random imputation procedure MRR_ϕ

Our imputation methods rely upon the mixture regression imputation model proposed by Haziza et al. (2014). As mentioned by these authors, the proposed methods could be extended to more general mixture regression models, for example to handle count data.

In practice, we may not be interested in the distribution function in itself, but rather in complex parameters such as quantiles. Establishing the theoretical properties of estimators of such parameters under the proposed imputation procedures is a challenging task, and is currently under investigation.

References

Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2010). Properties of design-based functional principal components analysis. *J. Stat. Plan.*

Infer., 140(1):75 – 91.

Cardot, H., Goga, C., Lardin, P., et al. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic journal of statistics*, 7:562–596.

Chauvet, G., Deville, J.-C., and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, 98(2):459–471.

Chauvet, G. and Do Paco, W. (2018). Exact balanced random imputation for sample survey data. *Computational Statistics & Data Analysis*, 128:1–16.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm for balanced sampling. *Computational Statistics*, 21(1):53–62.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey methodology*, 25(2):193–204.

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

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91(434):490–498.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Stat.*, 35:1491–1523.

Haziza, D. (1999). Imputation and inference in the presence of missing data. In Rao, C. and Pfeffermann, D., editors, *Handbook of Statistics, Sample Surveys: Theory Methods and Inference*, pages 215–246.

Haziza, D., Nambeu, C.-O., and Chauvet, G. (2014). Doubly robust imputation procedures for finite population means in the presence of a large number of zeros. *Canadian Journal of Statistics*, 42(4):650–669.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.*, 47:663–685.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *J. Am. Stat. Assoc.*, 77(377):89–96.

Kim, J. K. and Rao, J. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96(4):917–932.

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(445):254–265.

A Proof of Proposition 1

Lemma 1. *We have* $E\left\{\|\hat{G}_r - G\|^2\right\} = O(n^{-1})$.

Proof. *We can write* $\hat{G}_r - G = \left(\hat{G}_r - \tilde{G}_r\right) + \left(\tilde{G}_r - G\right)$, *where*

$$\tilde{G}_r = \frac{1}{N}\sum_{i\in s}\omega_i r_i \phi_i v_i^{-1} z_i z_i^\top. \tag{A.1}$$

With a proof similar to that of Lemma 2 in Chauvet and Do Paco (2018), we obtain $E\left\{\|\tilde{G}_r - G\|^2\right\} = O(n^{-1})$. *Also, we obtain from the assumptions:*

$$\left\|\hat{G}_r - \tilde{G}_r\right\| \leq \frac{C_6(C_9)^2C_{11}}{C_7}\|\hat{\gamma}_r - \gamma\|, \tag{A.2}$$

so that the result follows from Assumption (H6).

We can write $\hat{B}_{ar} - \beta = T_1 - T_2 + T_3$, where

$$\begin{aligned} T_1 &= \hat{G}_{ar}^{-1}\left\{\frac{1}{N}\sum_{i\in s}\omega_i r_i v_i^{-1} z_i(y_i - \phi_i z_i^\top \beta)\right\}, \\ T_2 &= \hat{G}_{ar}^{-1}\left\{\frac{1}{N}\sum_{i\in s}\omega_i r_i v_i^{-1}(\hat{\phi}_i - \phi_i) z_i z_i^\top\right\}\beta, \\ T_3 &= \hat{G}_{ar}^{-1}\left\{(\hat{G}_r - \hat{G}_{ar})1(\hat{G}_{ar} \neq \hat{G}_r)\right\}\beta. \end{aligned} \tag{A.3}$$

We have

$$\|T_1\|^2 \leq \frac{a^{-2}}{N^2}\sum_{i,j\in S} r_i r_j \omega_i \omega_j v_i^{-1} v_j^{-1} z_i^\top z_j (y_i - \phi_i z_i^\top \beta)(y_j - z_j^\top \beta). \tag{A.4}$$

Since the sampling design is non-informative and the response mechanism is unconfounded, we can write $E(\|T_1\|^2) = E_{pq}E_m(\|T_1\|^2)$ and

$$E(\|T_1\|^2) \leq E_{pq}\left[\frac{a^{-2}}{N^2}\sum_{i\in s} r_i \omega_i^2 v_i^{-2}\left\{\sigma^2\phi_i v_i + \phi_i(1-\phi_i)(z_i^\top\beta)^2\right\}\right] \tag{A.5}$$

and from the assumptions we obtain $E(\|T_1\|^2) = O(n^{-1})$. Also, we have

$$\|T_2\| \leq \frac{C_6(C_9)^2C_{11}}{aC_7}\|\hat{\gamma}_r - \gamma\|, \tag{A.6}$$

and from Assumption (H6) we obtain $E(\|T_2\|^2) = O(n^{-1})$. Finally, since $\|\hat{G}_r - \hat{G}_{ar}\|^2 \leq a^2$, we have

$$\begin{aligned} E(\|T_3\|^2) &\leq \|\beta\|^2 \times Pr(\hat{G}_{ar} \neq \hat{G}_r) \\ &\leq \frac{4\|\beta\|^2}{(\alpha_p - a)^2} E\left\{\|\hat{G}_r - G\|^2\right\}, \end{aligned} \tag{A.7}$$

where the second line in (A.7) follows from equation (B.21) in Chauvet and Do Paco (2018), and α_p is the largest eigenvalue of G given in equation (4.1). From Lemma 1, we have $E(\|T_3\|^2) = O(n^{-1})$, which completes the proof.

B Proof of Proposition 2

Lemma 2. *We have*

$$E\left\{(\bar{e}_r)^2\right\} = O(n^{-1}), \tag{B.1}$$

$$E\left\{\sigma_{er}^2\right\} = O(1). \tag{B.2}$$

Proof. *We consider equation (B.1) only. The proof of equation (B.2) is similar. We can rewrite* $\bar{e}_r = T_4 - T_5$*, with*

$$T_4 = \sum_{j \in s} \tilde{\omega}_j \eta_j r_j \epsilon_j \quad \textit{and} \quad T_5 = \left(\sum_{j \in s} \tilde{\omega}_j \eta_j r_j v_j^{-1/2} z_j\right)^\top (\hat{B}_{ar} - \beta). \tag{B.3}$$

It follows from the assumptions and from Proposition 1 that $E(T_5^2) = O(n^{-1})$.

We can rewrite $E(T_4^2) = \sigma^2 E(T_4')$*, with* $T_4' = \sum_{j \in s} \tilde{\omega}_j^2 \eta_j r_j$*. We note* $X =$

$\sum_{j\in s}\omega_j r_j\eta_j$, *and* $m_X = \sum_{j\in s}\omega_j p_j\phi_j$. *We can write* $T_4' = T_{41}' + T_{42}'$, *where* $T_{41}' = T_4' 1(X > m_X/2)$ *and* $T_{42}' = T_4 1(X \leq m_X/2)$. *From the assumptions, we have*

$$T_{41}' \leq \frac{4}{(C_4 C_4' C_5)^2} \times \frac{1}{N^2}\sum_{i\in s}\omega_i^2 p_i\phi_i, \tag{B.4}$$

which leads to $E(T_{41}') = o(n^{-1})$. *Also, since* $T_4' \leq 1$, *we have* $T_{42}' \leq 1(X \leq m_X/2)$ *and by using the Chebyshev inequality we obtain*

$$E(T_{42}'|s) \leq \frac{4}{(C_4 C_4' C_5)^2} \times \frac{1}{N^2}\sum_{i\in s}\omega_i^2 (p_i\phi_i)(1-p_i\phi_i), \tag{B.5}$$

which leads to $E(T_{42}') = o(n^{-1})$.

From the assumptions, we have $E\left[\left\{N^{-1}(\hat{t}_{y\pi} - t_y)\right\}^2\right] = O(n^{-1})$, so that it is sufficient to prove that $E\left[\left\{N^{-1}(\hat{t}_{yI} - \hat{t}_{y\pi})\right\}^2\right] = O(n^{-1})$. We have $N^{-1}(\hat{t}_{yI} - t_y) = T_6 + T_7 + T_8 + T_9$, with

$$\begin{aligned}
T_6 &= N^{-1}\sum_{i\in s} d_i(1-r_i)(y_i^* - \hat{\phi}_i z_i^\top \hat{B}_{ar}),\\
T_7 &= N^{-1}\sum_{i\in s} d_i(1-r_i)\hat{\phi}_i z_i^\top(\hat{B}_{ar} - \beta),\\
T_8 &= N^{-1}\sum_{i\in s} d_i(1-r_i)(\hat{\phi}_i - \phi_i) z_i^\top \beta,\\
T_9 &= N^{-1}\sum_{i\in s} d_i(1-r_i)(\phi_i z_i^\top \beta - y_i).
\end{aligned}$$

It readily follows from the assumptions, equation (**??**) and Proposition 1,

that $E(T_7^2) = o(1)$ and $E(T_8^2) = o(1)$. Also, since $E_m(T_9) = 0$, we obtain

$$E(T_9^2) = EV_m(T_9) = E\left[N^{-2}\sum_{i\in s} d_i^2(1-r_i)\left\{\sigma^2\phi_i v_i + \phi_i(1-\phi_i)(z_i^\top\beta)^2\right\}\right],$$

which is $O(n^{-1})$. Therefore, we only need to focus on T_6, for which we have

$$\begin{aligned} E_I(T_6^2) &= \left\{N^{-1}\sum_{i\in s} d_i(1-r_i)\hat{\phi}_i\sqrt{v_i}\right\}^2 (\bar{e}_r)^2 \\ &+ N^{-2}\sum_{i\in s} d_i^2(1-r_i)\left\{\hat{\phi}_i(1-\hat{\phi}_i)(z_i^\top\hat{B}_{ar} + \sqrt{v_i}\bar{e}_r)^2 + \hat{\phi}_i v_i\sigma_{er}^2\right\}. \end{aligned}$$

From Proposition 1 and Lemma 2, we obtain $E(T_6^2) = O(n^{-1})$.

C Proof of Proposition 3

From the assumptions, we have $E\left[\left\{\hat{F}_N(t) - F_N(t)\right\}^2\right] = O(n^{-1})$, so that it is sufficient to prove that $E\left[\left\{\hat{F}_I(t) - \hat{F}_N(t)\right\}^2\right] = o(1)$. We have $hat F_I(t) - \hat{F}_N(t) = T_{10} + T_{11} + T_{12}$, where

$$T_{10} = N^{-1}\sum_{i\in s} d_i(1-r_i)\left\{1(y_i^* \le t) - 1(y_i^{**} \le t)\right\}, \tag{C.1}$$

$$T_{11} = N^{-1}\sum_{i\in s} d_i(1-r_i)\left\{1(y_i^{**} \le t) - 1(\hat{y}_i \le t)\right\}, \tag{C.2}$$

$$T_{12} = N^{-1}\sum_{i\in s} d_i(1-r_i)\left\{1(\hat{y}_i \le t) - 1(y_i \le t)\right\}. \tag{C.3}$$

The values y_i^{**} and $\hat{y}_i$ are obtained as follows. We take

$$\hat{y}_i = \eta_i\left\{z_i^\top\beta + \sqrt{v_i}\hat{\epsilon}_i\right\}, \tag{C.4}$$

where $\hat{\epsilon}_i$ is selected with-replacement from the set $E'_r = \{\epsilon_j \ ; \ r_j = 1 \text{ and } \eta_j = 1\}$. We note $j(i)$ the donor selected for unit i, so that $\hat{\epsilon}_i = \epsilon_{j(i)}$. Also, we take

$$y_i^{**} \quad = \quad \eta_i \left\{ z_i^\top \hat{B}_{ar} + \sqrt{v_i} e_{g(i)} \right\} = \eta_i \left\{ z_i^\top \hat{B}_{ar} + \sqrt{v_i} \epsilon_i^* \right\}. \qquad \text{(C.5)}$$

We consider the term T_{10} first. We can write

$$1(y_i^* \leq t) - 1(y_i^{**} \leq t) \quad = \quad (\eta_i^* - \eta_i)\{1(\varepsilon_i^* \leq \hat{t}_i) - 1(t \geq 0)\}, \qquad \text{(C.6)}$$

with $\hat{t}_i = v_i^{-1/2}(t - z_i^\top \hat{B}_{ar}$. This leads to $(T_{10})^2 = T_{10,1} + T_{10,2}$, with

$$\begin{aligned}
T_{10,1} &= N^{-2} \sum_{i \in s} d_i^2 (1 - r_i)(\eta_i^* - \eta_i)^2 \{1(\varepsilon_i^* \leq \hat{t}_i) - 1(t \geq 0)\}^2, \\
T_{10,2} &= N^{-2} \sum_{i \neq j \in s} d_i (1 - r_i) d_j (1 - r_j)(\eta_i^* - \eta_i)(\eta_j^* - \eta_j) \times \\
&\qquad \{1(\varepsilon_i^* \leq \hat{t}_i) - 1(t \geq 0)\}\{1(\varepsilon_j^* \leq \hat{t}_j) - 1(t \geq 0)\}.
\end{aligned}$$

From the assumptions, $T_{10,1} = O(n^{-1})$. Also, since η_i^*, η_j^*, ε_i^* and ε_j^* are independent with respect to the imputation mechanism, we obtain successively

$$\begin{aligned}
E_I(T_{10,2}) &= N^{-2} \sum_{i \neq j \in s} d_i (1 - r_i) d_j (1 - r_j)(\hat{\phi}_i - \eta_i)(\hat{\phi}_j - \eta_j) \times \\
&\qquad \{\hat{F}_{\varepsilon_r}(\hat{t}_i) - 1(t \geq 0)\}\{\hat{F}_{\varepsilon_r}(\hat{t}_j) - 1(t \geq 0)\}
\end{aligned}$$

$$\begin{aligned}
E_m\{E_I(T_{10,2}) | \varepsilon_j, j \in s; \eta_g, g \in S_r\} &= N^{-2} \sum_{i \neq j \in s} d_i (1 - r_i) d_j (1 - r_j)(\hat{\phi}_i - \phi_i)(\hat{\phi}_j - \phi_j) \times \\
&\qquad \{\hat{F}_{\varepsilon_r}(\hat{t}_i) - 1(t \geq 0)\}\{\hat{F}_{\varepsilon_r}(\hat{t}_j) - 1(t \geq 0)\},
\end{aligned}$$

where $\hat{F}_{\varepsilon_r}(t) = \sum_{j\in s} \tilde{\omega}_j r_j \eta_j 1(e_j \leq t)$. This leads to

$$E(T_{10,2}) \;\leq\; \left(\frac{C_{11}}{C_1}\right)^2 E\left(\|\hat{\gamma}_r - \gamma\|^2\right) = o(1).$$

Consequently, $E(T_{10}^2) = o(1)$.

We now consider T_{11}, that we can write as

$$T_{11} \;=\; N^{-1} \sum_{i\in s} d_i (1-r_i)\eta_i \{1(\varepsilon_i^* \leq \hat{t}_i) - 1(\hat{\varepsilon}_i \leq t_i)\}$$

with $t_i = v_i^{-1/2}(t - z_i^\top \beta$, which leads to

$$\begin{aligned} E_I(|T_{11}|) &\leq N^{-1} \sum_{i\in s} d_i (1-r_i)\eta_i \sum_{j\in s} \tilde{\omega}_j r_j \eta_j \mid 1(e_j \leq \hat{t}_i) - 1(\varepsilon_j \leq t_i) \mid \\ &\leq N^{-1} \sum_{i\in s} d_i (1-r_i)\eta_i \sum_{j\in s} \tilde{\omega}_j r_j \eta_j \mid 1(\varepsilon_j \leq t_{ij}) - 1(\varepsilon_j \leq t_i) \mid \equiv T'_{11}, \end{aligned}$$

with

$$t_{ij} \;=\; t_i + \left(\frac{z_j}{\sqrt{v_j}} - \frac{z_i}{\sqrt{v_i}}\right)^\top (\hat{B}_{ar} - \beta).$$

Let us take some constant $\nu > 0$. Since the distribution function F_ε is absolutely continuous, there exists some τ_ν such that

$$|t-u| \leq \tau_\nu \;\;\Rightarrow\;\; |F_\varepsilon(t) - F_\varepsilon(u)| \leq \nu$$

We note $1_A = 1\left(\|\hat{B}_{ar} - \beta\| \geq 0.25\tau_\nu \sqrt{C_7}/C_9\right)$, and $1_B = 1 - 1_A$. We have $E\{T'_{11} 1(A)\} \leq (C_1)^{-1} E\{1(A)\}$, which is $o(1)$ from Proposition 1 and the

Chebyshev inequality. Also, we have

$$T'_{11}1(B) \leq N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i \sum_{j\in s}\tilde{\omega}_j r_j \eta_j 1\left(t_i - \frac{\tau_\nu}{2} \leq \varepsilon_j \leq t_i + \frac{\tau_\nu}{2}\right).$$

This leads to $E_m\{T'_{11}1(B)\} \leq (C_1)^{-1}\nu$, and since ν is arbitrary small, $E\{T'_{11}1(B)\} = o(1)$. Consequently, $E(|T_{11}|) = o(1)$.

Finally, we now consider T_{12} that we can write as

$$T_{12} = N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i \left\{1(\hat{\varepsilon}_i \leq t_i) - 1(\varepsilon_i \leq t_i)\right\}.$$

This successively leads to

$$\begin{aligned}
T_{12} &= N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i \left\{1(\hat{\varepsilon}_i \leq t_i) - 1(\varepsilon_i \leq t_i)\right\}, \qquad \text{(C.7)}\\
E_I(T_{12}) &= N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i \sum_{j\in s}\tilde{\omega}_j r_j \eta_j \left\{1(\varepsilon_j \leq t_i) - 1(\varepsilon_i \leq t_i)\right\},\\
E_m\{E_I(T_{12})|\eta_i, i\in s\} &= N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i \sum_{j\in s}\tilde{\omega}_j r_j \eta_j \left\{F_\varepsilon(t_i) - 1(F_\varepsilon(t_i)\right\} = 0,
\end{aligned}$$

and $E(T_{12}) = 0$, which gives

$$E\{(T_{12})^2\} = E_p E_q E_m V_I(T_{12}) + E_p E_q V_m E_I(T_{12}). \qquad \text{(C.8)}$$

We have $V_I(T_{12}) \leq C_1^{-1}n^{-1}$, so that the first term in the r.h.s. of (C.8) is $O(n^{-1})$. From the third line in equation (C.7), we obtain

$$V_m\{E_I(T_{12})\} = E_m V_m\{E_I(T_{12})|\eta_i, i\in s\}, \qquad \text{(C.9)}$$

and from the rewriting

$$E_I(T_{12}) = N^{-1}\sum_{j\in s}\tilde{\omega}_j r_j \eta_j \sum_{i\in s} d_i(1-r_i)\eta_i 1(\varepsilon_j \le t_i) - N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i 1(\varepsilon_i \le t_i),$$

we obtain

$$\begin{aligned}
V_m\{E_I(T_{12})|\eta_i, i\in s\} &= N^{-2}\sum_{j\in s}\tilde{\omega}_j^2 r_j\eta_j V_m\{\sum_{i\in s} d_i(1-r_i)\eta_i 1(\varepsilon_j \le\ t_i)|\eta_i, i\in s\} \\
&+ N^{-2}\sum_{i\in s} d_i^2(1-r_i)\eta_i F_\varepsilon(t_i)\{1-F_\varepsilon(t_i)\} \\
&= N^{-2}(\sum_{i\in s} d_i)^2 \sum_{j\in s}\tilde{\omega}_j^2 r_j\eta_j V_m\left\{\frac{\sum_{i\in s} d_i(1-r_i)\eta_i 1(\varepsilon_j \le\ t_i)}{\sum_{i\in s} d_i}\middle|\ \eta_i, i\in s\right\} \\
&+ N^{-2}\sum_{i\in s} d_i^2(1-r_i)\eta_i F_\varepsilon(t_i)\{1-F_\varepsilon(t_i)\} \\
&\le N^{-2}(\sum_{i\in s} d_i)^2 \sum_{j\in s}\tilde{\omega}_j^2\eta_j r_j + N^{-2}\sum_{i\in s} d_i^2. \\
&\le \frac{\sum_{j\in s}\tilde{\omega}_j^2\eta_j r_j + n^{-1}}{C_1^2}. \qquad \text{(C.10)}
\end{aligned}$$

From the proof of Lemma 2, we have $E(\sum_{j\in s}\tilde{\omega}_j^2\eta_j r_j) = O(n^{-1})$. From (C.9) and (C.10), we obtain that the second term in the r.h.s. of (C.8) is $O(n^{-1})$. Consequently, $E(T_{12}^2) = O(n^{-1})$. This completes the proof.