Exact balanced random imputation for sample survey data

Guillaume Chauvet, Wilfried Do Paco

Exact balanced random imputation for sample survey data

Guillaume Chauvet* Wilfried Do Paco†

August 2, 2017

Abstract

Surveys usually suffer from non-response, which decreases the effective sample size. Item non-response is typically handled by means of some form of random imputation if we wish to preserve the distribution of the imputed variable. This leads to an increased variability due to the imputation variance, and several approaches have been proposed for reducing this variability. Balanced imputation consists in selecting residuals at random at the imputation stage, in such a way that the imputation variance of the estimated total is eliminated or at least significantly reduced. In this work, we propose an implementation of balanced random imputation which enables to fully eliminate the imputation variance. Following the approach in Cardot et al. (2013), we consider a regularized imputed estimator of a total and of a distribution function, and we prove that they are consistent under the proposed imputation method. Some simulation results support our findings.

Key words: balanced imputation, cube method, distribution function, imputation mechanism, imputation model, mean-square consistency, regularized estimator

1 Introduction

Even if survey staff do their best in order to maximize response, it is unavoidable that surveys will suffer from some degree of non-response. This makes the effective sample size smaller, which results in an increase of the variance of estimators. More importantly, the respondents usually differ from the non-respondents with respect to the study variables. Therefore, unadjusted estimators will tend to be biased. In order to reduce the so-called non-response bias, it is therefore necessary to define estimation procedures accounting for non-response. Survey statisticians usually distinguish unit

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

nonresponse from item nonresponse. The former occurs when all variables are missing for some sampled unit, which may be due to a refusal to participate to the survey, or to the impossibility to contact the sampled unit, for example. The latter occurs when some variables, but not all, are missing for some sampled unit, which may be due to a refusal to answer to certain delicate questions in the survey, or to the length of the questionnaire, for example. Unit non-response is typically accounted for by reweighting estimators. Item non-response is typically handled by means of some form of imputation, which consists in replacing missing values with artificial values in order to reduce the bias and possibly control the variance due to non-response. In this paper, we are interested in imputation procedures to treat item non-response.

Imputation methods may be classified into two broad classes: deterministic and random. Deterministic imputation methods yield a fixed imputed value given the sample. For example, deterministic regression imputation consists in using a regression model to predict the missing value for a non-respondent, making use of auxiliary information available for the whole sample, including non-respondents. Deterministic regression imputation leads to an approximately unbiased estimator of the total if the regression model is correctly specified. However, deterministic imputation tends to distort the distribution of the imputed variable, and some form of random imputation is typically used if we wish to preserve the distribution of the imputed variable. Random imputation methods are closely related to deterministic imputation methods, except that a random term is added to the prediction in order to mimic as closely as possible the relationship between the variable of interest and the explanatory variables. The main drawback of random imputation methods is that they lead to estimators with increased variability due to the imputation variance. In some cases, the contribution of the imputation variance to the global variance may be large, resulting in potentially inefficient estimators.

The data collected for a given survey are typically used to estimate a variety of parameters. The survey is often primarily designed to estimate totals over a given population of interest. For example, this may be the total biomass of living vegetation for a region in case of a forest survey (Gregoire and Valentine, 2008, page 3), or the total of revenues and expenses inside categories of firms in case of business surveys (Haziza et al., 2016). On the other hand, secondary analysts may be interested in estimating more complex parameters, some of them being directly linked to the population distribution function like the quantiles (Boistard et al., 2016). Therefore, a same variable of interest is habitually used to estimate several parameters. A random imputation method is needed if we wish to preserve the distribution of this variable, but the imputation mechanism should be chosen so that the imputation variance is kept as small as possible for the estimation of the total of the variable.

In the literature, three general approaches for reducing the imputation variance have been considered. The fractional imputation approach consists of replacing each missing value with $M \geq 2$ imputed values selected randomly, and assigning a weight to each imputed value (Kalton and Kish, 1981, 1984; Fay, 1996; Kim and Fuller, 2004; Fuller and Kim, 2005). It can be shown that the imputation variance decreases as M increases. The second approach consists of first imputing the missing values using a standard random imputation method, and then adjusting the imputed values in such a way that the imputation variance is eliminated; see Chen et al. (2000). The third approach that we study consists of selecting residuals at random in such a way that the imputation variance is eliminated (Kalton and Kish, 1981, 1984; Deville, 2006; Chauvet et al., 2011; Hasler and Tillé, 2014).

In this work, we propose an implementation of balanced random imputation that makes it possible to fully eliminate the imputation variance for the estimation of a total. Also, we propose regularized imputed estimators of a total and of the distribution function, following the approach in Cardot et al. (2013), and we establish their consistency. The paper is organized as follows. Our notations in case of full response are defined in Section 2, and the principles of balanced sampling are briefly reminded. The imputation model is presented in Section 3, along with the imputed estimators of the total and of the distribution function. The regularized estimator of the model parameter is also introduced. In Section 4, we describe the proposed exact balanced random imputation method. We give an illustration on a small dataset, and we prove the consistency of the imputed estimator of the total and of the imputed estimator of the distribution function. The results of a simulation study are presented in Section 5. We draw some conclusions in Section 6. The proofs are deferred to the Appendix.

2 Finite population framework

2.1 Notation

We consider a finite population U of size N with some variable of interest y. We are interested in estimating some finite population parameter such as the total $t_y = \sum_{k \in U} y_k$ or the population distribution function

$$F_N(t) \quad = \quad N^{-1} \sum_{k \in U} 1(y_k \leq t) \tag{2.1}$$

with $1(\cdot)$ the indicator function.

In order to study the asymptotic properties of the sampling designs and estimators that we treat below, we consider the asymptotic framework of Isaki and Fuller (1982). We assume that the population U belongs to a nested sequence $\{U_t\}$ of finite populations with increasing sizes N_t, and that

the population vector of values $y_{Ut} = (y_{1t}, \ldots, y_{Nt})^\top$ belongs to a sequence $\{y_{Ut}\}$ of N_t-vectors. For simplicity, the index t will be suppressed in what follows and all limiting processes will be taken as $t \to \infty$.

A random sample S is selected in U by means of some sampling design $p(\cdot)$, which is a probability distribution defined over the subsets of the population U. That is, we have

$$p(s) \geq 0 \text{ for any } s \subset U \text{ and } \sum_{s \subset U} p(s) = 1. \tag{2.2}$$

We assume that the sampling design is of fixed size n, which means that a subset s has a probability of selection equal to zero if this subset is not of size n. We note I_k for the sample membership indicator, equal to 1 if the unit k is selected in the sample S and to 0 otherwise. We note $I_U = (I_1, \ldots, I_N)^\top$ for the vector of sample indicators. Since the sampling design is of fixed size n, we have

$$\sum_{k \in U} I_k \quad = \quad n. \tag{2.3}$$

2.2 Inclusion probabilities

The probability for unit k to be included in the sample is denoted as π_k. We note $\pi_U = (\pi_1, \ldots, \pi_N)^\top$ for the vector of inclusion probabilities. All the inclusion probabilities are assumed to be non-negative, i.e. there is no coverage bias in the population. Since $\pi_k = E_p(I_k)$, with E_p the expectation with respect to the sampling design $p(\cdot)$, we obtain from equation (2.3) that

$$\sum_{k \in U} \pi_k \quad = \quad n. \tag{2.4}$$

We also denote by π_{kl} the probability that units k and l are selected jointly in the sample.

In case of equal inclusion probabilities, we have $\pi_k = n/N$ for any unit $k \in U$. This occurs for example if the sample is selected by means of simple random sampling without replacement. Another customary choice consists in using inclusion probabilities proportional to some auxiliary non-negative variable z_{1k}, known for any unit $k \in U$. This leads to a so-called probability proportional to size (π-ps) sampling design, which is used in some business surveys (e.g., Ohlsson, 1998). In such case, we obtain from (2.4) that

$$\pi_k \quad = \quad n \frac{z_{1k}}{\sum_{l \in U} z_{1l}}. \tag{2.5}$$

If some units exhibit a large value for the auxiliary variable z_{1k}, equation (2.5) may lead to inclusion probabilities greater than 1. In this case, these inclusion probabilities are set to 1, which means that

the corresponding units are selected in the sample with certainty, and the inclusion probabilities for the remaining units are computed by means of equation (2.5) restricted to the remaining units (see Tillé, 2006, page 18).

In a situation of full response, a design-unbiased estimator for t_y is the Horvitz-Thompson estimator

$$\hat{t}_{y\pi} = \sum_{k\in U} d_k I_k y_k = \sum_{k\in S} d_k y_k \tag{2.6}$$

with $d_k = \pi_k^{-1}$ the sampling weight, and an approximately unbiased estimator for $F_N(t)$ is

$$\hat{F}_N(t) = \frac{1}{\hat{N}} \sum_{k\in S} d_k 1(y_k \leq t) \quad \text{with} \quad \hat{N} = \sum_{k\in S} d_k. \tag{2.7}$$

2.3 Balanced sampling

Suppose that a q-vector x_k of auxiliary variables is known at the design stage for any unit $k \in U$. A sampling design $p(\cdot)$ is said to be balanced on x_k if the vector I_U of sample indicators is such that

$$\sum_{k\in U} \frac{x_k}{\pi_k} I_k = \sum_{k\in U} x_k. \tag{2.8}$$

In other words, the sampling design is balanced on x_k if for any possible sample, the Horvitz-Thompson estimator of the total of the auxiliary variables exactly matches the true total.

Deville and Tillé (2004) introduced a sampling design called the cube method, which enables to select balanced samples, or approximately balanced samples if an exact balancing is not feasible. The cube method proceeds through a random walk from the vector π_U of inclusion probabilities to the vector I_U of sample indicators. This random walk proceeds in two steps. At the end of the first one called the flight phase (see Appendix A), we obtain a random vector $\tilde{I}_U = (\tilde{I}_1, \ldots, \tilde{I}_N)^\top$ such that

$$\sum_{k\in U} \frac{x_k}{\pi_k} \tilde{I}_k = \sum_{k\in U} x_k, \tag{2.9}$$

and such that $\tilde{I}_k = 0$ if the unit k is definitely rejected from the sample, $\tilde{I}_k = 1$ if the unit k is definitely selected in the sample, and $0 < \tilde{I}_k < 1$ if the decision for unit k remains pending. From equation (2.9), the balancing is exactly respected at the end of the flight phase, but we do not obtain a sample per se since the decision remains pending for some units. From Proposition 1 in Deville and Tillé (2004), it can be shown that the number of such units is no greater than q, the number of auxiliary variables. A second step called the landing phase is then applied on the set of remaining units, in order to end the sampling while ensuring that the balancing equations (4.2) remain approximately satisfied. This leads to the vector of sample indicators I_U.

3 Imputed estimators

In a situation of item non-response, the variable y is observed for a subsample of units only. We note r_k for a response indicator for unit k, and ϕ_k for the response probability of unit k. We note n_r the number of responding units, and n_m the number of missing units. We assume that the units respond independently. In case of simple imputation, an artificial value y_k^* is used to replace the missing y_k and leads to the imputed version of the HT-estimator

$$\hat{t}_{yI} = \sum_{k\in S} d_k r_k y_k + \sum_{k\in S} d_k (1-r_k) y_k^*, \tag{3.1}$$

and to the imputed version of the estimated distribution function

$$\hat{F}_I(t) = \frac{1}{\hat{N}} \left\{ \sum_{k\in S} d_k r_k 1(y_k \leq t) + \sum_{k\in S} d_k (1-r_k) 1(y_k^* \leq t) \right\}. \tag{3.2}$$

3.1 Imputation model

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

$$m: \; y_k = f(z_k, \beta) + v_k^{1/2} \epsilon_k, \tag{3.3}$$

where $f(\cdot)$ is a given function, z_k is a K-vector of auxiliary variables available at the imputation stage for all $k \in S$, β is a K-vector of unknown parameters and v_k is a known constant. The ϵ_k are assumed to be independent and identically distributed random variables with mean 0 and variance σ^2, with a common distribution function denoted as $F_\epsilon(\cdot)$ and where σ is an unknown parameter. The model (3.3) is often called an imputation model (e.g., Särndal, 1992; Chauvet et al, 2011).

In practice, most imputation techniques which are used in surveys are motivated by a particular case of the imputation model m when $f(z_k, \beta) = z_k^\top \beta$. This is true for mean imputation, where a missing value is replaced by the mean of respondents; for hot-deck imputation, where a missing value is replaced by randomly selecting an observed value among the respondents; for regression imputation, where a missing value is replaced by a prediction by regression, to which a random noise is added in case of random regression imputation (see Haziza, 2009). In order to simplify the presentation, we therefore focus on the linear case $f(z_k, \beta) = z_k^\top \beta$ in the remainder of the paper, which leads to the so-called regression imputation model

$$m: \; y_k = z_k^\top \beta + v_k^{1/2} \epsilon_k. \tag{3.4}$$

A particular case of model (3.4) called the ratio imputation model is presented in Section 4.3 for illustration.

In this paper, inference will be made with respect to the joint distribution induced by the imputation model, the sampling design and the non-response mechanism, which is known as the Imputation Model approach (IM). We assume that the sampling design is non-informative (Särndal et al, 1992, p. 33; Pfeffermann, 2009), namely that the vector of sample membership indicators $I_U \equiv (I_1, \ldots, I_N)^\top$ is independent of $\epsilon_U \equiv (\epsilon_1, \ldots, \epsilon_N)^\top$, conditionally on a set of design variables $x_U \equiv (x_1, \ldots, x_N)^\top$. We do not need an explicit modeling of the non-response mechanism, unlike the Non-response Model approach (see Haziza, 2009). However, the data are assumed to be missing at random (see Rubin, 1976, 1983) in the sense that the vector of response indicators $r_U \equiv (r_1, \ldots, r_N)^\top$ is related to a set of auxiliary variables $z_U \equiv (z_1, \ldots, z_N)^\top$ known for any unit k in S, but the vector r_U is independent of the vector y_U, conditionally on z_U.

3.2 Imputation mechanism

Mimicking the imputation model (3.4), the imputed value is

$$y_k^* = z_k^\top \hat{B} + v_k^{1/2} \epsilon_k^*, \tag{3.5}$$

with $\hat{B}$ some estimator of β. Using $\epsilon_k^* = 0$ in (3.5), we obtain deterministic regression imputation which leads to an approximately unbiased estimation for the total t_y but not for the distribution function $F_N(\cdot)$. Therefore, we focus in the rest of the paper on random regression imputation, where the imputed values in (3.5) are obtained by generating the random residuals ϵ_k^* randomly. For each unit k for which y_k is missing, we select a donor, which is a responding unit for which the value of the variable of interest is used to fill-in the missing value for unit k. More precisely, the random residuals ϵ_k^* are selected from the set of observed residuals

$$E_r = \{e_l;\ r_l = 1\} \quad \text{where} \quad e_l = \frac{y_l - z_l^\top \hat{B}}{v_l^{1/2}}. \tag{3.6}$$

The residual e_l is attributed to the non-respondent k with the probability

$$Pr(\epsilon_k^* = e_l) = \tilde{\omega}_l \quad \text{where} \quad \tilde{\omega}_l = \frac{\omega_k}{\sum_{l \in S} \omega_l r_l}, \tag{3.7}$$

where ω_l is an imputation weight attached to unit l. We assume that these imputation weights do not depend on ϵ_U, I_U or r_U. Alternatively, the residuals ϵ_k^* could be generated from a given parametric distribution.

A possible estimator for the unknown parameter β is

$$\hat{B}_r = \hat{G}_r^{-1} \left(\frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k y_k \right) \quad \text{with} \quad \hat{G}_r = \frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k z_k^\top. \tag{3.8}$$

Since a matrix $\hat{G}_r$ close to singularity can lead to unstable estimators, we follow the approach proposed in Cardot et al. (2013) and we introduce a regularized version of $\hat{B}_r$. We first write

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

where $\eta_{1r} \geq \ldots \geq \eta_{pr}$ are the non-negative eigenvalues of $\hat{G}_r$, with $u_{1r}, \ldots, u_{pr}$ the corresponding orthonormal eigenvectors. For a given $a > 0$, the regularized version of $\hat{G}_r$ as defined in Cardot et al. (2013) is then

$$\hat{G}_{ar} = \sum_{j=1}^{p} \max(\eta_{jr}, a) u_{jr} u_{jr}^\top, \tag{3.10}$$

which is an invertible matrix with

$$\|\hat{G}_{ar}^{-1}\| \quad \leq \quad a^{-1}, \tag{3.11}$$

where $\|\cdot\|$ stands for the spectral norm. This leads to the regularized estimator of the parameter β

$$\hat{B}_{ar} = \hat{G}_{ar}^{-1} \left(\frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k y_k \right). \tag{3.12}$$

In the rest of the paper, we use this regularized estimator in (3.5) to generate the imputed values, and in (3.6) to define the observed residuals.

3.3 Consistency of the regularized estimator

In order to study the asymptotic properties of the estimators that we treat below, we consider the following regularity assumptions:

H1: There exists some constant $C_1, C_2 > 0$ such that $C_1 \leq N n^{-1} \pi_k \leq C_2$ for any $k \in U$.

H2: There exists some constant C_3 such that $\sup_{k \neq l \in U} \left(n \left| 1 - \frac{\pi_{kl}}{\pi_k \pi_l} \right| \right) \leq C_3$.

H3: There exists some constant $C_4 > 0$ such that $C_4 \leq \min_{k \in U} \phi_k$.

H4: There exists some constants $C_5, C_6 > 0$ such that $C_5 \leq N^{-1} n \omega_k \leq C_6$ for any $k \in U$.

H5: There exists some constants $C_7, C_8 > 0$ such that $C_7 \leq v_k \leq C_8$ for any $k \in U$. There exists some constant C_9 such that $\|z_k\| \leq C_9$ for any $k \in U$. Also, the matrix

$$G \quad = \quad \frac{1}{N} \sum_{k \in U} \pi_k \phi_k \omega_k v_k^{-1} z_k z_k^\top \tag{3.13}$$

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

Assumptions (H1) and (H2) are related to the sampling design. In case of sampling with equal probabilities, we have $\pi_k = n/N$ and Assumption (H1) is automatically fulfilled with $C_1 = C_2 = 1$. The assumption (H1) means that in case of sampling with unequal probabilities, the inclusion probabilities do not depart much from that obtained when sampling with equal probabilities. In Assumption (H2), the quantity $\sup_{k \neq l \in U} \left(n \left| 1 - \frac{\pi_{kl}}{\pi_k \pi_l} \right| \right)$ is a measure of dependency in the selection of units. This quantity is equal to zero when the units in the population are selected independently, which is known as Poisson sampling (Fuller, 2009, p. 13). This assumption is satisfied for many sampling designs like stratified simple random sampling or rejective sampling, see for example Cardot et al. (2013). Both assumptions (H1) and (H2) are classical in survey sampling.

Assumption (H3) is related to the response mechanism. It is assumed that the response probabilities are bounded below from zero, i.e. that all units in the population have a strictly positive probability to answer the survey. Assumption (H4) is related to the imputation mechanism. If $\omega_k = N/n$ for any unit $k \in U$, all the responding units have the same probability of being selected to fill-in a missing value. It is thus assumed in (H4) that when selecting the random residuals, no extreme imputation weight will dominate the others. The assumption (H5) is in particular related to the choice of the regularizing parameter a, and is needed to guarantee the point-wise convergence of the estimator $\hat{B}_{ar}$ for the regression coefficient. A similar assumption is considered in Cardot et al. (2013).

Proposition 1. Assume that the imputation model (3.3) holds, and that assumptions (H1)-(H5) hold. Then:

$$E\left\{ \|\hat{B}_{ar} - \beta\|^2 \right\} \quad = \quad O(n^{-1}). \tag{3.14}$$

4 Balanced random imputation

4.1 Motivation

In practice, a survey serves multiple purposes. On the one hand, the survey designer is interested in estimating aggregate parameters such as totals, and the variance of total estimates needs to be kept as small as possible. On the other hand, the survey data are used by secondary analysts who are interested in other parameters of interest which may be related to the distribution of the imputed variable, like quantiles. Therefore, a random imputation method is used to fill-in the missing values in order to preserve the distribution of the imputed variable. In the same time, the random residuals need to be generated in such a way that the variance of $\hat{t}_{yI}$ does not suffer from the variance imputation.

The drawback of random regression imputation lies indeed in an additional variability for the estimation of t_y, called the imputation variance. The imputed estimator of the total may be

written as

$$\hat{t}_{yI} = \sum_{k\in S} d_k r_k y_k + \sum_{k\in S} d_k(1-r_k)(z_k^\top \hat{B}_{ar}) + \sum_{k\in S} d_k(1-r_k)(v_k^{1/2}\epsilon_k^*). \quad (4.1)$$

The imputation variance is due to the third term on the right-hand side only. This imputation variance is completely eliminated if the random residuals are selected so that

$$\begin{aligned}\sum_{k\in S} d_k(1-r_k)v_k^{1/2}\epsilon_k^* &= E_I\left\{\sum_{k\in S} d_k(1-r_k)v_k^{1/2}\epsilon_k^*\right\} \\ &= \sum_{k\in S}\left\{d_k(1-r_k)v_k^{1/2}\right\}\bar{e}_r \quad (4.2)\end{aligned}$$

with $\bar{e}_r = \sum_{j\in S}\tilde{\omega}_j r_j e_j$.

4.2 Exact balanced random imputation

Our goal is therefore to generate the random residuals in such a way that equation (4.2) holds. A natural idea would be to select the random residuals ϵ_k^* directly with replacement from the set E_r of observed residuals. Unfortunately, to the best of our knowledge, there does not exist any general with-replacement sampling design which enables to select the random residuals such that equation (4.2) holds. In order to be able to use the cube method presented in Section 2.3, we follow the approach in Chauvet et al. (2011) and proceed in 4 steps:

1. We build a population U^* of $n_m \times n_r$ cells, each row being associated to one non-respondent and each column being associated to one respondent.

2. To each cell $(k,l) \in U^*$, we associate a selection probability $\psi_{kl} = \tilde{\omega}_l$ (see equation 3.7) and a value $x_{kl}^0 = d_k v_k^{1/2}\psi_{kl} e_l$.

3. We apply the flight phase of the cube method on population U^*, with inclusion probabilities $\psi_{U^*} = (\psi_{11},\ldots,\psi_{n_m n_r})^\top$, by balancing on the variable x_{kl}^0. From equation (2.9), we obtain at the end of the flight phase a random vector $\tilde{I}_{U^*} = (\tilde{I}_{11},\ldots,\tilde{I}_{n_m n_r})^\top$ such that

$$\sum_{(k,l)\in U^*} \frac{x_{kl}^0}{\psi_{kl}}\tilde{I}_{kl} = \sum_{(k,l)\in U^*} x_{kl}^0. \quad (4.3)$$

4. For any non-respondent k, the imputed residual is

$$\epsilon_k^* = \sum_{l\in S_r} \tilde{I}_{kl} e_l. \quad (4.4)$$

Table 1: Values of the guess of the amount of money, of the true amount of money and of observed residuals for a simple random sample of $n = 10$ persons

Unit k	1	2	3	4	5	6	7	8	9	10
z_{1k}	8.35	1.5	10	0.6	7.5	7.95	0.95	4.4	1	0.5
y_k	8.75	2.55	9	1.1	7.5	5				
e_k	0.30	0.93	-0.14	0.69	0.15	-0.89				

From equations (4.3) and (4.4), it is easily shown that equation (4.2) holds. Therefore, the imputation variance of $\hat{t}_{yI}$ is completely eliminated under the proposed imputation procedure. This is an advantage as compared to the balanced imputation procedure in Chauvet et al. (2011), where the balancing constraint (4.3) was not exactly respected, and the imputation variance was therefore not fully eliminated. A drawback of the proposed method is that a missing residual is not necessarily replaced by an observed estimated residual, but may be replaced by a weighted mean of observed estimated residuals, which may result in a bias in the estimation of the distribution function. However, by adding n_m balancing variables in the imputation procedure, we ensure that ϵ_k^* is an observed residual for at least $n_m - 1$ units. The additional n_m-vector of balancing variables is

$$x \quad = \quad (x^1, \ldots, x^i, \ldots, x^{n_m})^\top, \tag{4.5}$$

with $x_{kl}^i = \psi_{kl} 1(k = i)$ for the cell (k, l), see Chauvet et al. (2011). We prove in Proposition 3 that using this additional set of balancing variables in the imputation process enables to preserve the distribution of the imputed variable.

4.3 An illustration of the proposed method

We illustrate the proposed imputation method on a small sample, based on an example presented in Thompson (2002, p. 70). In a population U of $N = 53$ persons in a lecture theater, each person k is asked to write down a guess of the amount of money (variable z_{1k}) he/she is carrying. A simple random sample of $n = 10$ persons is then selected, and each person is asked to write down the exact amount of money (variable y_k) he/she is carrying. For this illustration, we consider that the variable y_k is missing for 4 persons in the sample. The data are presented in Table 1.

We consider so-called ratio imputation, which is obtained from (3.5) in case of a single auxiliary variable ($z_k = z_{1k}$) and with $v_k = z_{1k}$. This leads to the ratio imputation model

$$m: \; y_k = \beta z_{1k} + z_{1k}^{1/2} \epsilon_k, \tag{4.6}$$

which is currently used in business surveys. We use equal imputation weights $\omega_k = 1$. This leads

Table 2: A first example of balanced random imputation

Non-respondent	Observed residuals						Imputed	Imputed
k	0.30	0.93	-0.14	0.69	0.15	-0.89	residual ϵ_k^*	value y_k^*
7	0	0	0	1	0	0	0.69	1.57
8	0.61	0	0	0	0	0.39	-0.17	3.80
9	1	0	0	0	0	0	0.30	1.24
10	1	0	0	0	0	0	0.30	0.68

to

$$y_k^* = \hat{B}_r z_{1k} + z_{1k}^{1/2}\epsilon_k^* \quad \text{where} \quad \hat{B}_r = \frac{\sum_{k\in S} r_k y_k}{\sum_{k\in S} r_k z_{1k}}. \tag{4.7}$$

In this example, we obtain $\hat{B}_r = 0.94$. The observed residuals e_k for respondents are given in the last line of Table 1. We have

$$\sum_{k\in S} d_k(1-r_k)v_k^{1/2}\bar{e}_r \quad = \quad 4.38. \tag{4.8}$$

In order to impute the missing values, we create a table of 4×6 cells with one row for each non-respondent and one column for each observed residual. We then draw a sample of 4 cells by means of the flight phase of the cube method. The result is presented in Table 2. The 4-th cell on row 1 is selected, which means that we take $\epsilon_7^\star = e_4 = 0.69$. Similarly, the 1-th cell on row 3 is selected so that $\epsilon_9^\star = e_1$, and the 1-th cell on row 4 is selected so that we take $\epsilon_{10}^\star = e_1$. On row 2, we obtain $\tilde{I}_{21} = 0.61$ and $\tilde{I}_{26}^\star = 0.39$, so that we take $\epsilon_9^\star = 0.61 * e_1 + 0.39 * e_6 = -0.17$. With these imputed residuals, we obtain

$$\sum_{k\in S} d_k(1-r_k)v_k^{1/2}\epsilon_k^* \quad = \quad 4.38 \tag{4.9}$$

so that from equation (4.8), the balancing equation is exactly respected. We present in Table 3 another possible set of imputed residuals. It can be shown that equation (4.8) holds, so that the balancing equation is satisfied and the imputation variance for the total is eliminated.

4.4 Properties of balanced random imputation

It is shown in Proposition 2 below that the imputed estimator of the total is mean-square consistent for the true total. Also, we prove in Proposition 3 that the imputed distribution function under the proposed exact balanced imputation procedure is consistent for the population distribution function.

Proposition 2. Assume that the imputation model (3.3) holds, and that assumptions (H1)-(H5) hold. Assume that the exact balanced imputation procedure is used. Then:

$$E[\{N^{-1}(\hat{t}_{yI} - t_y)\}^2] \quad = \quad O(n^{-1}). \tag{4.10}$$

Table 3: A second example of balanced random imputation

Non-respondent	Observed residuals						Imputed	Imputed
k	0.30	0.93	-0.14	0.69	0.15	-0.89	residual ϵ_k^*	value y_k^*
7	0	0.83	0	0	0	0.17	0.62	1.50
8	1	0	0	0	0	0	0.30	4.78
9	0	0	0	0	0	1	-0.89	0.06
10	0	0	0	1	0	0	0.69	0.96

Proposition 3. We assume that assumptions (H1)-(H5) hold. Assume that F_ϵ is absolutely continuous. Assume that the exact balanced imputation procedure is used, and that the balancing variables in (4.5) are added. Then:

$$E\left|\hat{F}_I(t) - F_N(t)\right| \quad = \quad o(1). \tag{4.11}$$

5 Simulation study

We conducted a simulation study to test the performance of several imputation methods in terms of relative bias and relative efficiency. We first generated 2 finite populations of size $N = 10,000$, each containing one study variable y and one auxiliary variable z_1. In each population, the variable z_1 was first generated from a Gamma distribution with shape and scale parameters equal to 2 and 5, respectively. Then, given the z_1-values, the y-values were generated according to the model $y_k = \beta\ z_{1k} + z_{1k}^{1/2}\ \epsilon_k$ presented in equation (4.6). The parameter β was set to 1 and the ϵ_k were generated according to a normal distribution with mean 0 and variance σ^2, whose value 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.

We were interested in estimating two parameters: the population total of the y-values, t_y and the finite population distribution function, $F_N(t)$ for $t = t_\alpha$, where t_α is the α-th population quantile. We considered $\alpha = 0.25$ and 0.50 in the simulation. From each population, we selected $1,000$ samples of size $n = 100$ by means of rejective sampling also called conditional Poisson sampling (e.g., Hajek, 1964) with inclusion probabilities, π_k, proportional to z_k. That is, we have $\pi_k = nz_k/t_z$, where $t_z = \sum_{k\in U} z_k$. Then, in each generated sample, nonresponse to item y was generated according to two nonresponse mechanisms which are described below:

MCAR: uniform response mechanism, where all the units in U have the same probability of response ϕ_0. We used $\phi_0 = 0.5$ and $\phi_0 = 0.75$.

MAR: the probability ϕ_k of response attached to unit k is defined as

$$\log\left(\frac{\phi_k}{1-\phi_k}\right) = \lambda_0 + \lambda_1 z_{1k}, \tag{5.1}$$

where the parameters λ_0 and λ_1 were chosen so that the average $\bar{\phi}$ of the ϕ_k's was approximately equal to 0.5, or approximately equal to 0.75.

In each sample containing respondents and nonrespondents, imputation was performed according to three methods, all motivated by the imputation model (4.6). The imputed values are given by

$$y_k^* = \hat{B}_r z_{1k} + z_{1k}^{1/2}\epsilon_k^*. \tag{5.2}$$

For deterministic ratio imputation (DRI), the imputed values are given by (5.2) with $\epsilon_k^* = 0$ for all k. The imputed values for random ratio imputation (RRI) are given by (5.2), where the residuals ϵ_k^* are selected independently and with replacement. The imputed values for exact balanced ratio imputation (EBRI) are given by (5.2) where the residuals ϵ_k^* are selected so that the balancing constraint (4.2) is exactly satisfied.

Then, we computed the imputed estimator of t_y given by (3.1), and the imputed estimator of $F_N(t)$ given by (3.2). As a measure of the bias of an estimator $\hat{\theta}_I$ of a parameter θ, we used the Monte Carlo percent relative bias

$$\mathrm{RB}(\hat{\theta}_I) = \frac{E_{MC}(\hat{\theta}_I) - \theta}{\theta} \times 100, \tag{5.3}$$

where $E_{MC}(\hat{\theta}_I) = \sum_{r=1}^{1000} \hat{\theta}_I^{(r)}/1000$, and $\hat{\theta}_I^{(r)}$ denotes the estimator $\hat{\theta}_I$ in the r-th sample, $r = 1, \ldots, 1000$. As a measure of variability of $\hat{\theta}_I$, we used the Monte Carlo mean square error

$$\mathrm{MSE}(\hat{\theta}_I) = \frac{1}{1000}\sum_{r=1}^{1000}(\hat{\theta}_I^{(r)} - \theta)^2. \tag{5.4}$$

Let $\hat{\theta}_I^{(DRI)}$, $\hat{\theta}_I^{(RRI)}$, and $\hat{\theta}_I^{(EBRI)}$ denote the estimator $\hat{\theta}_I$ under deterministic ratio imputation, random ratio imputation and exact balanced ratio imputation, respectively. In order to compare the relative efficiency of the imputed estimators, using $\hat{\theta}_I^{(RRI)}$ as the reference, we used

$$\mathrm{RE} = \frac{\mathrm{MSE}(\hat{\theta}_I^{(.)})}{\mathrm{MSE}(\hat{\theta}_I^{(RRI)})}. \tag{5.5}$$

Monte Carlo measures for $\hat{F}_I(t)$ were obtained from (5.3)-(5.5) by replacing $\hat{\theta}_I$ with $\hat{F}_I(t)$ and θ_N with $F_N(t)$.

Table 4 shows the values of relative bias and relative efficiency corresponding to the imputed estimator $\hat{t}_{yI}$. It is clear from Table 4 that $\hat{t}_{yI}$ was approximately unbiased in all the scenarios, as

Table 4: Monte Carlo percent relative bias of the imputed estimator and relative efficiency

		DRI	RRI	EBRI	DRI	RRI	EBRI
		MCAR					
		$\phi_0 = 0.5$			$\phi_0 = 0.75$		
Population 1	RB	0.47	0.50	0.47	0.30	0.33	0.30
	RE	0.79	1	0.79	0.79	1	0.79
Population 2	RB	0.17	0.26	0.17	0.16	0.25	0.16
	RE	0.79	1	0.79	0.79	1	0.79
		MAR					
		$\bar{\phi} = 0.5$			$\bar{\phi} = 0.75$		
Population 1	RB	0.28	0.30	0.28	0.45	0.62	0.45
	RE	0.69	1	0.69	0.72	1	0.72
Population 2	RB	0.02	-0.06	0.02	-0.18	-0.14	-0.18
	RE	0.70	1	0.70	0.74	1	0.74

expected. In terms of relative efficiency, results showed that DRI and EBRI lead to the smallest mean square error for the estimation of a total. This result is not surprising since the imputation variance is identically equal to zero for both imputation methods. We note that DRI and EBRI were particulary efficient in the MAR case.

We now turn to the distribution function, $F_N(t)$. Table 5 shows the relative bias and relative efficiency corresponding to the imputed estimator $\hat{F}_I(t)$. As expected, the estimators under deterministic ratio imputation were considerably biased, and the absolute relative bias can be as high as 42.4%. In terms of relative bias, both RRI and EBRI showed almost no bias, except for $t_{0.25}$ in the case of balanced imputation. These results can be explained by the fact that both imputation methods succeeded in preserving the distribution of the study variable y. Also, we note that the imputed estimator $\hat{F}_I(t)$ under exact balanced ratio imputation was more efficient than the corresponding estimator under random ratio imputation in all the scenarios with a value of relative efficiency varying from 0.89 to 1.00. The lower values of RE were obtained in the MAR case.

6 Final remarks

In this paper, we considered estimation under item non-response. We proposed an exact balanced random imputation procedure, where the imputation variance is completely eliminated for the estimation of a total. We also proved that the proposed imputation procedure leads to mean-square consistent estimators for a total and for a distribution function.

Table 5: Monte Carlo percent relative bias of the imputed estimator of the distribution function and relative efficiency

			DRI	RRI	EBRI	DRI	RRI	EBRI
			MCAR					
	α		$\phi_0 = 0.5$			$\phi_0 = 0.75$		
Population 1	0.25	RB	-41.3	-1.6	-2.7	-31.3	-1.1	-2.0
		RE	2.03	1	0.94	1.66	1	0.94
	0.50	RB	-4.7	-1.3	-0.9	-3.6	-0.7	-0.6
		RE	1.22	1	0.98	1.13	1	0.97
Population 2	0.25	RB	-26.7	-0.7	-1.4	-22.2	-0.5	-1.1
		RE	1.45	1	0.93	1.34	1	0.94
	0.50	RB	-2. 7	-0.3	-0.1	-2.1	-0.1	0.1
		RE	1.09	1	0.97	1.07	1	0.97
			MAR					
	α		$\phi_0 = 0.5$			$\phi_0 = 0.75$		
Population 1	0.25	RB	-42.4	-0.3	-0.9	-24.2	-2.5	-3.5
		RE	2.11	1	0.89	1.37	1	0.90
	0.50	RB	2.1	-0.0	0.2	5.6	-1.5	-0.2
		RE	1.18	1	0.95	1.12	1	0.96
Population 2	0.25	RB	-15.4	0.6	0.4	-3.4	1.3	2.0
		RE	1.17	1	0.93	1.00	1	0.93
	0.50	RB	0.2	0.1	-0.3	2.6	-0.7	-0.2
		RE	1.09	1	1.00	1.02	1	0.96

We have not considered the problem of variance estimation in the context of the proposed balanced random imputation. Variance estimation for the imputed estimator of the total is fairly straightforward, since the imputed estimator is identical to that under deterministic regression imputation. Variance estimation for the imputed distribution function is currently under investigation.

When studying relationships between study variables, Shao and Wang (2002) proposed a joint random regression imputation procedure which succeeds in preserving the relationship between these variables, and a balanced version of their procedure was proposed by Chauvet and Haziza (2012). Extending the exact balanced random procedure to this situation is a matter for further research.

References

Bhatia, R. (1997). Matrix analysis. *Springer-Verlag.*

Boistard, H., and Chauvet, G., and Haziza, D. (2016). Doubly robust inference for the distribution function in the presence of missing survey data. *Scandinavian Journal of Statistics*, **43**, 683-699.

Cardot, H., and Goga, C., and Lardin, P. (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., and Deville, J.C., and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, **98(2)**, 459-471.

Chauvet, G. and Haziza, D. (2012). Fully efficient estimation of coefficients of correlation in the presence of imputed survey data. *Canadian Journal of Statistics*, **40(1)**, 124-149.

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

Chen, J., and Rao, J. N. K., and Sitter, R. R. (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica*, **10(4)**, 1153-1169.

Deville (2006). Random imputation using balanced sampling. *Presentation to the Joint Statistical Meeting of the American Statistical Association, Seattle, USA*.

Deville, J. C., and Särndal, C. E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, **10**, 381-394.

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.

Fuller, W. A. (2009). Sampling statistics. *Wiley*.

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

Gregoire, T.G., and Valentine, H.T. (2008). Sampling strategies for natural resources and the environment. *Chapman & Hall: Boca Raton*.

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

Hasler, C., and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics and Data Analysis*, **74**, 81-94.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of Statistics*, **29**, 215-246.

Haziza, D., and Nambeu, C., and Chauvet, G. (2016). 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.

Isaki, C. T., and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77(377)**, 89-96.

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

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics-Theory and Methods*, **13(16)**, 1919-1939.

Kim, J. K., and Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, **91(3)**, 559-578.

Ohlsson, E. (1998). Sequential Poisson Sampling. *Journal of Official Statistics*, **14(2)**, 149-162.

Pfeffermann, D. (2009). Inference under informative sampling. *Handbook of Statistics*, **29(B)**, 455-487.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63(3)**, 581-592.

Rubin, D. B. (1983). Conceptual issues in the presence of nonresponse. *Incomplete data in sample surveys*, **2**, 123-142.

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

Särndal, C. E., and Swensson, B., and Wretman, J. (1992). Model assisted survey sampling. *Springer*.

Shao, J., and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation. *Journal of the American Statistical Association*, **97(458)**, 544-552.

Thompson, W. A. (2002). Sampling. *Wiley*.

Tillé, Y. (2006). Sampling algorithms. *Springer: New-York*.

A Flight phase of the cube method (Tillé, 2006, p. 160)

We define the balancing matrix as $A = (x_1/\pi_1, \ldots, x_N/\pi_N)$. We initialize with $\pi_U(0) = \pi_U$. Next, at time $t = 0, \ldots, T$, repeat the three following steps.

Step 1: Let $E(t) = F(t) \cap \mathrm{Ker} A$, where

$$F(t) \quad = \quad \{v \in \mathbb{R}^N : v_k = 0 \text{ if } \pi_k(t) \text{ is an integer}\},$$

with $\pi_U(t) = (\pi_1(t), \ldots, \pi_N(t))^\top$. If $E(t) \neq \{0\}$, generate any vector $v(t) \neq 0$ in $E(t)$, random or not.

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

$$0 \leq \pi_U(t) + \lambda_1(t)v(t) \leq 1 \quad \text{and} \quad 0 \leq \pi_U(t) - \lambda_2(t)v(t) \leq 1,$$

where the inequalities are interpreted element-wise. Note that $\lambda_1^*(t) > 0$ and $\lambda_2^*(t) > 0$.

Step 3: Take $\pi_U(t+1) = \pi_U(t) + \delta_U(t)$, where

$$\delta_U(t) \quad = \quad \begin{cases} \lambda_1^*(t)v(t) & \text{with probability } \frac{\lambda_2^*(t)}{\lambda_1^*(t)+\lambda_2^*(t)}, \\ -\lambda_2^*(t)v(t) & \text{with probability } \frac{\lambda_1^*(t)}{\lambda_1^*(t)+\lambda_2^*(t)}. \end{cases}$$

The flight phase ends at time T, when it is no longer possible to find a non-null vector in $E(T)$. The random vector obtained at the end of the flight phase is $\tilde{I}_U = \pi(T)$.

B Proof of Proposition 1

Lemma B.1. *We have*

$$E_p\left[(\hat{t}_{y\pi} - t_y)^2\right] \quad \leq \quad \left(\sup_{k\neq l\in U} n\left|1 - \frac{\pi_{kl}}{\pi_k\pi_l}\right|\right) \sum_{k\in U} \pi_k \left(\frac{y_k}{\pi_k} - \frac{t_y}{n}\right)^2, \tag{B.1}$$

$$E_p\left[\{\hat{F}_N(t) - F_N(t)\}^2\right] \quad \leq \quad \left(\frac{4}{N^2}\right)\left(\sup_{k\neq l\in U} n\left|1 - \frac{\pi_{kl}}{\pi_k\pi_l}\right|\right) \sum_{k\in U} \frac{1}{\pi_k}. \tag{B.2}$$

Proof . The proof is standard, and is therefore omitted.

Lemma B.2. *We have:*

$$E(\|\hat{G}_r - G\|^2) \quad = \quad O(n^{-1}). \tag{B.3}$$

Proof . We note $\|\cdot\|_F$ for the Frobenius norm. Using the fact that the spectral norm is smaller than the Frobenius norm, we have

$$\begin{aligned} E(\|\hat{G}_r - G\|^2) &\leq E(\|\hat{G}_r - G\|_F^2) \\ &= E\left[\frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\omega_k v_k^{-1}(I_k r_k - \pi_k\phi_k)\omega_l v_l^{-1}(I_l r_l - \pi_l\phi_l) tr(z_k z_k^\top z_l z_l^\top)\right] \\ &= T_3 + T_4 \end{aligned} \tag{B.4}$$

with

$$T_3 = \frac{1}{N^2}\sum_{k\in U}\omega_k^2 v_k^{-2}\pi_k\phi_k(1-\pi_k\phi_k)tr(z_k z_k^\top z_k z_k^\top), \tag{B.5}$$

$$T_4 = \frac{1}{N^2}\sum_{k\neq l\in U}\omega_k v_k^{-1}\omega_l v_l^{-1}(\pi_{kl} - \pi_k\pi_l)\phi_k\phi_l tr(z_k z_k^\top z_l z_l^\top). \tag{B.6}$$

Since $tr(z_k z_k^\top z_k z_k^\top) = \|z_k\|^4$, we obtain from assumptions (H1), (H4) and (H5)

$$T_3 \leq \left(\frac{(C_6)^2 C_2}{n(C_7)^2}\right)\left(\frac{1}{N}\sum_{k\in U}\|z_k\|^4\right), \tag{B.7}$$

which is $O(n^{-1})$ from assumption (H5). Also, we have

$$T_4 \leq \frac{1}{N^2}\left(\sup_{k\neq l\in U}\left|1 - \frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\sum_{k\neq l\in U}\omega_k v_k^{-1}\omega_l v_l^{-1}\pi_k\pi_l\phi_k\phi_l tr(z_k z_k^\top z_l z_l^\top). \tag{B.8}$$

Since

$$tr(z_k z_k^\top z_l z_l^\top) = (z_k^\top z_l)^2 \leq \|z_k\|^2\|z_l\|^2, \tag{B.9}$$

we obtain

$$\begin{aligned} T_4 &\leq \frac{1}{N^2}\left(\sup_{k\neq l\in U}\left|1 - \frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\left(\sum_{k\in U}\omega_k v_k^{-1}\pi_k\phi_k\|z_k\|^2\right)^2 \\ &\leq \frac{(C_2)^2(C_6)^2}{n(C_7)^2}\left(\sup_{k\neq l\in U} n\left|1 - \frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\left(\frac{1}{N}\sum_{k\in U}\|z_k\|^2\right)^2, \end{aligned} \tag{B.10}$$

which is $O(n^{-1})$ from assumptions (H2) and (H5). This completes the proof of Lemma B.2.

We now consider the proof of Proposition 1. We can write

$$\hat{B}_{ar} - \beta = T_5 + T_6, \tag{B.11}$$

where

$$T_5 = \hat{G}_{ar}^{-1}\left\{\frac{1}{N}\sum_{k\in S} r_k\omega_k v_k^{-1} z_k(y_k - z_k^\top\beta)\right\}, \tag{B.12}$$

$$T_6 = \hat{G}_{ar}^{-1}\left\{(\hat{G}_r - \hat{G}_{ar})1(\hat{G}_{ar}\neq\hat{G}_r)\right\}\beta. \tag{B.13}$$

We first consider the term T_5. We have:

$$\begin{aligned} \|T_5\|^2 &\leq \|\hat{G}_{ar}^{-1}\|^2 \times \left\| \frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k (y_k - z_k^\top \beta) \right\|^2 \\ &\leq a^{-2} \times \frac{1}{N^2} \sum_{k,l \in S} r_k r_l \omega_k \omega_l v_k^{-1} v_l^{-1} z_k^\top z_l (y_k - z_k^\top \beta)(y_l - z_l^\top \beta), \end{aligned} \tag{B.14}$$

where the second line in (B.14) follows from (3.11). Since the sampling design is non-informative and the response mechanism is unconfounded, we can write

$$E(\|T_5\|^2) = E_{pq} E_m(\|T_5\|^2), \tag{B.15}$$

where $E_{pq}(\cdot)$ stands for the expectation with respect to the sampling design and the response mechanism, and $E_m(\cdot)$ stands for the expectation with respect to the imputation model conditionally on I_U and r_U. From (B.14), (B.15), and from the assumptions on the imputation model (3.3), we obtain

$$\begin{aligned} E(\|T_5\|^2) &\leq E_{pq} \left\{ \sigma^2 a^{-2} \times \frac{1}{N^2} \sum_{k \in S} r_k \omega_k^2 v_k^{-1} z_k^\top z_k \right\} \\ &\leq \left(\frac{\sigma^2 a^{-2} (C_6)^2 C_2}{n C_7} \right) \left(\frac{1}{N} \sum_{k \in U} \|z_k\|^2 \right) \end{aligned} \tag{B.16}$$

where the second line in (B.16) follows from assumptions (H1), (H4) and (H5). From assumption (H5), this leads to $E(\|T_5\|^2) = O(n^{-1})$.

We now consider the term T_6, by following the same lines as in Lemma A.1 of Cardot et al. (2013). We have:

$$\begin{aligned} \|T_6\|^2 &\leq \|\hat{G}_{ar}^{-1}\|^2 \times \left\| (\hat{G}_r - \hat{G}_{ar}) 1(\hat{G}_{ar} \neq \hat{G}_r) \right\|^2 \times \|\beta\|^2 \\ &\leq a^{-2} \|\beta\|^2 \times \left\| (\hat{G}_r - \hat{G}_{ar}) 1(\hat{G}_{ar} \neq \hat{G}_r) \right\|^2. \end{aligned} \tag{B.17}$$

Since $\|\hat{G}_{ar} - \hat{G}_r\|^2 \leq a^2$, we obtain

$$E(\|T_6\|^2) \leq \|\beta\|^2 \times Pr(\hat{G}_{ar} \neq \hat{G}_r). \tag{B.18}$$

We write

$$G = \sum_{j=1}^{p} \eta_j u_j u_j^\top, \tag{B.19}$$

where $\eta_1 \geq \ldots \geq \eta_p$ are the non-negative eigenvalues of G, with $u_1, \ldots, u_p$ the corresponding orthonormal eigenvectors. We have

$$\begin{aligned} Pr(\hat{G}_{ar} \neq \hat{G}_r) &= Pr(\eta_{pr} \neq a) \\ &\leq Pr\left(|\eta_{pr} - \eta_p| \geq \frac{|\eta_p - a|}{2}\right) \\ &\leq \frac{4}{(\eta_p - a)^2} E(|\eta_{pr} - \eta_p|^2) \qquad \text{(B.20)} \\ &\leq \frac{4}{(\eta_p - a)^2} E\|\hat{G}_r - G\|^2 \qquad \text{(B.21)} \end{aligned}$$

where equation (B.20) follows from the Chebyshev inequality, and equation (B.21) follows from the fact that the eigenvalue map is Lipschitzian for symmetric matrices (see Bhatia (1997), chapter 3, and Cardot et al. (2013), p. 580). From (B.18) and (B.21), and using Lemma B.2, we obtain $E(\|T_6\|^2) = O(n^{-1})$. This completes the proof.

C Proof of Proposition 2

From equation (B.1), we obtain under Assumptions (H1), (H2), (H5) and under the model assumptions that

$$E[\{N^{-1}(\hat{t}_{y\pi} - t_y)\}^2] = O(n^{-1}). \qquad \text{(C.1)}$$

It is therefore sufficient to prove that

$$E[\{N^{-1}(\hat{t}_{yI} - \hat{t}_{y\pi})\}^2] = O(n^{-1}). \qquad \text{(C.2)}$$

We can write $N^{-1}(\hat{t}_{yI} - \hat{t}_{y\pi}) = T_7 - T_8$, where

$$\begin{aligned} T_7 &= N^{-1} \sum_{k \in S} d_k (1 - r_k) z_k^\top (\hat{B}_{ar} - \beta), \qquad \text{(C.3)} \\ T_8 &= \sigma N^{-1} \sum_{k \in S} d_k (1 - r_k) v_k^{1/2} \epsilon_k. \qquad \text{(C.4)} \end{aligned}$$

We have

$$\begin{aligned} |T_7|^2 &\leq N^{-2} \left\| \sum_{k \in S} d_k (1 - r_k) z_k \right\|^2 \times \left\| \hat{B}_{ar} - \beta \right\|^2 \\ &\leq (C_9/C_1)^2 \left\| \hat{B}_{ar} - \beta \right\|^2, \qquad \text{(C.5)} \end{aligned}$$

where the second line in (C.5) follows from Assuptions (H1) and (H5). From Proposition 1, we obtain $E(|T_7|^2) = O(n^{-1})$.

We now turn to T_8. We have $E_m(T_8) = 0$, so that

$$E(T_8^2) = V(T_8) = E_p E_q V_m(T_8). \tag{C.6}$$

Also, we have

$$V_m(T_8) \quad = \quad \sigma^2 N^{-2} \sum_{k \in S} d_k^2 (1 - r_k) v_k, \tag{C.7}$$

which is $O(n^{-1})$ from Assumptions (H1) and (H5). This completes the proof.

D Proof of Proposition 3

We can write

$$\hat{F}_I(t) - F_N(t) \quad = \quad \left\{ \hat{F}_I(t) - \tilde{F}_I(t) \right\} + \left\{ \tilde{F}_I - F_N(t) \right\}, \tag{D.1}$$

where

$$\tilde{F}_I(t) \quad = \quad \hat{N}^{-1} \left\{ \sum_{k \in S} d_k r_k 1(y_k \leq t) + \sum_{k \in S} d_k r_k 1(y_k^{**} \leq t) \right\}, \tag{D.2}$$

and $y_k^{**} = z_k^\top \hat{B}_{ar} + v_k^{1/2} \epsilon_k^{**}$ is the imputed value under the balanced random imputation procedure of Chauvet et al. (2011).

Since the number of units such that $0 < \tilde{I}_{kl} < 1$ at the end of the flight phase is bounded, we have $y_k^* = y_k^{**}$ for all units in S_m but a bounded number of units. Therefore, there exists some constant C such that

$$\begin{aligned} |\hat{F}_I(t) - \tilde{F}_I(t)| \quad &\leq \quad \hat{N}^{-1} \times C \sup_{k \in S} d_k \\ &\leq \quad \frac{C \; C_2}{C_1 \; n}, \end{aligned} \tag{D.3}$$

where the second line in (D.3) follows from Assumption (H1). Therefore,

$$E|\hat{F}_I(t) - \tilde{F}_I(t)| \quad = \quad O(n^{-1}). \tag{D.4}$$

It follows from the proof of Theorem 2 in Chauvet et al. (2011) that

$$E|\tilde{F}_I(t) - F_N(t)| \quad = \quad o(1). \tag{D.5}$$

From (D.1), the proof is complete.