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 12, 2016

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 donors or residuals at random at the imputation stage, in such a way that the imputation variance 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 for some parameters. 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 their mean square consistency. 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

Surveys usually suffer from non-response, which decreases the effective sample size. 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. Imputation methods may be classified into two broad classes: deterministic and random. Unlike random imputation methods, if the imputation process is repeated, deterministic methods yield a fixed imputed value given the sample. Some form of random imputation is typically used if we wish to preserve the distribution of the imputed variable, but leads to estimators with increased variability due to the imputation variance. In some cases, the contribution of

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

the imputation variance is appreciable resulting in potentially inefficient estimators.

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 donors or 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 which enables 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 mean square consistency. The paper is organized as follows. Our notations in case of full response are defined in Section 2. 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, and we prove the mean square 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

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 fixed-size sampling design $p(\cdot)$. The probability for unit k to be included in the sample is denoted as π_k, and is assumed to be non-negative. In a situation of full response, a design-unbiased estimator for t_y is the Horvitz-Thompson estimator

$$\hat{t}_{y\pi} = \sum_{k \in S} d_k y_k \tag{2.2}$$

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

$$\hat{F}_N(t) = \frac{1}{\hat{N}} \sum_{k \in S} d_k 1(y_k \leq t) \quad \text{with} \quad \hat{N} = \sum_{k \in S} d_k. \tag{2.3}$$

Proposition 1. We have

$$E_p\left[(\hat{t}_{y\pi} - t_y)^2\right] \leq \left(\sup_{k \neq l \in U} n \left|1 - \frac{\pi_{kl}}{\pi_k \pi_l}\right|\right) \sum_{k \in U} \pi_k \left(\frac{y_k}{\pi_k} - \frac{t_y}{n}\right)^2, \tag{2.4}$$

$$E_p\left[\{\hat{F}_N(t) - F_N(t)\}^2\right] \leq \left(\frac{4}{N^2}\right) \left(\sup_{k \neq l \in U} n \left|1 - \frac{\pi_{kl}}{\pi_k \pi_l}\right|\right) \sum_{k \in U} \frac{1}{\pi_k}, \tag{2.5}$$

with $E_p(\cdot)$ the expectation with respect to the sampling design $p(\cdot)$.

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 p_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

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

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

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

where 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, σ is an unknown parameter 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 1, and with a common distribution function denoted as $F_\epsilon(\cdot)$. The model (3.3) is often called an imputation model (e.g., Särndal, 1992).

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). An explicit modeling of the non-response mechanism is not needed, unlike the Non-response Model approach (see Haziza, 2009), but the data are assumed to be missing at random as defined by Rubin (1976). 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 on $\epsilon_U \equiv (\epsilon_1, \ldots, \epsilon_N)^\top$, conditionally on a set of design variables $x_U \equiv (x_1, \ldots, x_N)^\top$. Also, we assume that the non-response mechanism is unconfounded (Deville and Särndal, 1994; Rubin, 1983), namely that the vector of response indicators $r_U \equiv (r_1, \ldots, r_N)^\top$ is independent on ϵ_U, conditionally on the set of auxiliary variables $z_U \equiv (z_1, \ldots, z_N)^\top$.

Mimicking the imputation model, the imputed value is

$$y_k^* = z_k^\top \hat{B} + \hat{\sigma} v_k^{1/2} \epsilon_k^*, \tag{3.4}$$

with $\hat{B}$ and $\hat{\sigma}$ some estimators of β and σ. Using $\epsilon_k^* = 0$ in (3.4), 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 (3.4) are obtained by selecting the random residuals ϵ_k^* from the set of observed residuals

$$E_r = \{e_l;\ r_l = 1\} \quad \text{where} \quad e_l = \frac{y_l - z_l^\top \hat{B}}{\hat{\sigma} v_l^{1/2}}. \tag{3.5}$$

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

$$Pr(\epsilon_k^* = e_l) = \tilde{\omega}_l \quad \text{where} \quad \tilde{\omega}_l = \frac{\omega_k}{\sum_{l \in S} \omega_l r_l}, \tag{3.6}$$

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

$$\hat{B}_r = \hat{G}_r^{-1} \left(\frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k y_k \right) \quad \text{with} \quad \hat{G}_r = \frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k z_k^\top. \tag{3.7}$$

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

$$\hat{G}_r = \sum_{j=1}^{p} \eta_{jr} u_{jr} u_{jr}^\top, \tag{3.8}$$

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

$$\hat{G}_{ar} = \sum_{j=1}^{p} \max(\eta_{jr}, a) u_{jr} u_{jr}^\top, \tag{3.9}$$

which is an invertible matrix with

$$\|\hat{G}_{ar}^{-1}\| \quad \leq \quad a^{-1}, \tag{3.10}$$

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

$$\hat{B}_{ar} = \hat{G}_{ar}^{-1} \left(\frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k y_k \right). \tag{3.11}$$

In the rest of the paper, we use this regularized estimator in (3.4) to generate the imputed values, and in (3.5) to define the observed residuals.

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 $C_4 \leq \min_{k \in U} p_k$.

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

$$G \quad = \quad \frac{1}{N} \sum_{k \in U} \pi_k p_k \omega_k v_k^{-1} z_k z_k^\top \tag{3.12}$$

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

The assumptions (H1) and (H2) are classical in survey sampling. It is assumed in (H3) that the response probabilities are bounded below from zero. It is assumed in (H4) that no extreme imputation weight dominates the others. The assumption (H5) is in particular related to the choice of the regularizing parameter a; similar assumptions are also considered in Cardot et al. (2013).

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

$$E\left\{\|\hat{B}_{ar}-\beta\|^2\right\} \;=\; O(n^{-1}). \tag{3.13}$$

4 Balanced random imputation

Performing random regression imputation amounts to select the random residuals ϵ_k^* with replacement from the set of observed residuals E_r. As pointed out by Chauvet et al. (2011), this may be performed by selecting a without-replacement sample S^* from a population U^* of $n_m \times n_r$ cells, where selecting the cell (k,l) means that the donor l is attributed to the non-respondent k, i.e. $\epsilon_k^* = \tilde{e}_l$. So as to respect exactly the drawing probabilities in (3.6), each cell $(k,l) \in U^*$ is given the probability of selection $\psi_{kl} = \tilde{\omega}_l$. So as to select one donor exactly for each non-respondent, we can proceed by stratifying the population U^* by rows, and by using a one unit per stratum stratified design for the selection of S^*.

One drawback of random regression imputation lies in an additional variability called the imputation variance. The imputed estimator of the total may be written as

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

and 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

$$\begin{aligned}\sum_{k\in S} d_k(1-r_k)v_k^{1/2}\epsilon_k^* &= E_I\left\{\sum_{k\in S} d_k(1-r_k)v_k^{1/2}\epsilon_k^*\right\}\\ &= \sum_{k\in S} d_k(1-r_k)v_k^{1/2}\bar{e}_r \text{ with } \bar{e}_r = \sum_{j\in S}\tilde{\omega}_j r_j e_j.\end{aligned} \tag{4.2}$$

The sampling design used for the selection of S^* may be chosen so that equation (4.2) holds, at least approximately (see Kalton and Kish, 1981, 1984; Deville, 2006; Chauvet et al., 2011). It may be shown that equation (4.2) is equivalent to the first balancing equation

$$\sum_{(k,l)\in U^*} \frac{x_{kl}^0}{\psi_{kl}} I_{kl}^* = \sum_{(k,l)\in U^*} x_{kl}^0, \tag{4.3}$$

with $x_{kl}^0 = d_k v_k^{1/2} \psi_{kl} e_l$ for the cell (k,l), and with $I_{U^*} = (I_{11}^*, \ldots, I_{n_m n_r}^*)^\top$ the vector of sample membership indicators in S^*. Also, selecting exactly one donor for each non-respondent is equivalent to the second set of balancing equations

$$\sum_{(k,l)\in U^*} \frac{x_{kl}}{\psi_{kl}} I_{kl}^* = \sum_{(k,l)\in U^*} x_{kl}, \tag{4.4}$$

with $x = (x^1, \ldots, x^i, \ldots, x^{n_m})^\top$ and $x_{kl}^i = \psi_{kl} 1(k=i)$ for the cell (k,l). Therefore, Chauvet et al. (2011) proposed to select a sample S^* balanced on variables $\tilde{x}^\top = (x^0, x^\top)$ by means of the cube method (Deville and Tillé, 2004). Other constraints may be added for the selection of S^* if it is desired to eliminate the imputation variance for other parameters.

The cube method for balanced random imputation proposed by Chauvet et al. (2011) proceeds in two steps: a flight phase, at the end of which an exact balancing is maintained, and a landing phase in which the balancing equations are partly relaxed until the complete sample S^* is obtained. The flight phase (Deville and Tillé, 2004; Chauvet and Tillé, 2006; Tillé, 2006) proceeds through a random walk from the vector of probabilities $\psi_{U^*} = (\psi_{11}, \ldots, \psi_{n_m n_r})^\top$ to a random vector $\psi_{U^*}^* = (\psi_{11}^*, \ldots, \psi_{n_m n_r}^*)^\top$ such that $\psi_{kl}^* = 0$ is the cell (k,l) is definitely rejected, $\psi_{kl}^* = 1$ is the cell (k,l) is definitely selected, and $0 < \psi_{kl}^* < 1$ if the decision for (k,l) remains pending at the end of the flight phase. The balancing equations are perfectly respected at the end of the flight phase, in the sense that

$$\sum_{(k,l)\in U^*} \frac{\tilde{x}_{kl}}{\psi_{kl}} \psi_{kl}^* = \sum_{(k,l)\in U^*} \tilde{x}_{kl}. \tag{4.5}$$

The landing phase enables to end the sampling, either by successively relaxing the balancing equations or by means of an enumerative algorithm on the remaining units. At the end of the landing phase, the final vector of sample selection indicators I_{U^*} is obtained. The random residual for unit $k \in S_m$ is then

$$\epsilon_k^{**} = \sum_{l\in S_r} I_{kl}^* e_l. \tag{4.6}$$

With this balanced random imputation procedure, the set of balancing equations (4.4) is exactly respected. As a result, for any unit $k \in S_m$ one unit l exactly in S_r is such that $I_{kl}^* = 1$. Denoting by $l(k)$ this unit, we can therefore rewrite (4.6) as

$$\epsilon_k^{**} = e_{l(k)}. \tag{4.7}$$

The missing residual is therefore replaced with an observed estimated residual. A drawback of this balanced imputation procedure is that, due to the landing phase, the imputation variance is not

completely eliminated. At the end of the landing phase, we have

$$\sum_{(k,l)\in U^*} \frac{\tilde{x}_{kl}}{\psi_{kl}} I^*_{kl} \simeq \sum_{(k,l)\in U^*} \tilde{x}_{kl}. \tag{4.8}$$

Therefore, the balancing equation (4.3) is only approximately respected and the imputation variance is not completely eliminated. Our proposal is to follow the same approach, but to use the result of the flight phase only for the matter of imputation. That is, the random residuals are obtained as

$$\epsilon^*_k = \sum_{l\in S_r} \psi^*_{kl} e_l. \tag{4.9}$$

From equation (4.5), this choice enables to fulfill equation (4.2) exactly. The imputed estimator of the total is therefore exactly the same as under deterministic regression imputation, and the imputation variance is completely eliminated. It is shown in Proposition 3 below that the imputed estimator of the total is mean-square consistent for the true total.

Proposition 3. Assume that the imputation model (3.3) holds, and that assumptions (H1)-(H5) hold. Then:

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

A drawback of the proposed method is that a missing residual is not necessarily replaced by an observed estimated residual, but may be replaced with a weighted mean of observed estimated residuals. However, it can be shown that with the proposed balanced random imputation procedure, ϵ^*_k is an observed residual for at least $n_m - 1$ units. Even if the vector $\tilde{x}$ includes additional balancing constraints, e.g. if it is desired to eliminate the imputation variance for other parameters, then the number of units such that $0 < \psi^*_{kl} < 1$ at the end of the flight phase is bounded (Hasler and Tillé, 2014). Therefore, a missing residual ϵ_k is replaced with an observed estimated residual, for all units in S_m but perhaps a bounded number of units which are replaced with weighted means of observed residuals. We prove in Proposition 4 below that the imputed distribution function under the proposed exact balanced imputation procedure is mean-square consistent for the population distribution function.

Proposition 4. We assume that assumptions (H1)-(H5) hold. Also, assume that $\hat{\sigma}$ is a consistent estimator of σ, and that F_ϵ is absolutely continuous. Then:

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

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. In each population, the variable z was first generated from a Gamma distribution with shape and scale parameters equal to 2 and 5, respectively. Then, given the z-values, the y-values were generated according to the model $y_k = \beta\ z_k + z_k^{1/2}\ \eta_k$. The parameter β was set to 1 and the η_i 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 was generated according to an uniform nonresponse mechanism with a probability of response p_0. We considered $p_0 = 0.5$ and 0.75 in the simulation. In each sample containing respondents and nonrespondents, imputation was performed according to three methods: deterministic ratio imputation, random ratio imputation and exact balanced ratio imputation. All three methods are motivated by the imputation model (3.3) with z_k scalar and $v_k = z_k$. For deterministic ratio imputation (DRI), the imputed values are given by (3.4) with $\epsilon_k^* = 0$ for all k. The imputed values for random ratio imputation (RRI) are given by (3.4), where the residuals ϵ_k^* are selected independently and with replacement, while the imputed values for exact balanced ratio imputation (EBRI) are given by (3.4) where the residuals ϵ_k^* are selected so that the balancing constraint (4.3) 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

$$\mathrm{RB}(\hat{\theta}_I) = \frac{E_{MC}(\hat{\theta}_I) - \theta}{\theta} \times 100, \tag{5.1}$$

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

$$\mathrm{MSE}(\hat{\theta}_I) = \frac{1}{1000} \sum_{r=1}^{1000} (\hat{\theta}_I^{(r)} - \theta)^2. \tag{5.2}$$

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

$$\mathrm{RE} = \frac{\mathrm{MSE}(\hat{\theta}_I^{(.)})}{\mathrm{MSE}(\hat{\theta}_I^{(RRI)})}. \tag{5.3}$$

Table 1: Monte Carlo percent relative bias of the imputed estimator and relative efficiency

		DRI	RRI	EBRI	DRI	RRI	EBRI
		$p_0 = 0.5$			$p_0 = 0.75$		
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

Table 2: Monte Carlo percent relative bias of the imputed estimator of the distribution function and relative efficiency

			DRI	RRI	EBRI	DRI	RRI	EBRI
	α		$p_0 = 0.5$			$p_0 = 0.75$		
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

Monte Carlo measures for $\hat{F}_I(t)$ were obtained from (5.1)-(5.3) by replacing $\hat{\theta}_I$ with $\hat{F}_I(t)$ and θ_N with $F_N(t)$.

Table 1 shows the values of relative bias and relative efficiency corresponding to the imputed estimator $\hat{t}_{yI}$. It is clear from Table 1 that $\hat{t}_{yI}$ was approximately unbiased in all the scenarios, as 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 now turn to the distribution function, $F_N(t)$. Table 2 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, with absolute relative bias ranging from 2.7% to 41.3%. 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 0.93 to 0.98.

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.

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*.

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

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.

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.

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.

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

A Proof of Proposition 1

We first consider equation (2.4). We have

$$\begin{aligned} E_p\left[(\hat{t}_{y\pi}-t_y)^2\right] &= \frac{1}{2}\sum_{k\neq l\in U}(\pi_k\pi_l-\pi_{kl})\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2 \qquad (A.1)\\ &\leq \left(\sup_{k\neq l\in U}\left|1-\frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\times\frac{1}{2}\sum_{k\neq l\in U}\pi_k\pi_l\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2. \end{aligned}$$

From the identity

$$\frac{1}{2}\sum_{k\neq l\in U}\pi_k\pi_l\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2 = n\sum_{k\in U}\pi_k\left(\frac{y_k}{\pi_k}-\frac{t_y}{n}\right)^2, \qquad (A.2)$$

we obtain (2.4). We now turn to equation (2.5). We have $\hat{F}_N(t)-F_N(t)=T_1+T_2$ with

$$T_1 = \left\{\frac{N-\hat{N}}{N}\right\}\left\{\frac{1}{\hat{N}}\sum_{k\in S}d_k1(y_k\leq t)\right\}, \qquad (A.3)$$

$$T_2 = \frac{1}{N}\left\{\sum_{k\in S}d_k1(y_k\leq t)-\sum_{k\in U}1(y_k\leq t)\right\}. \qquad (A.4)$$

We have $T_1^2\leq N^{-2}(\hat{N}-N)^2$, and by replacing y_k with 1 in equation (2.4), we obtain

$$E_p(T_1^2) \leq \left(\frac{1}{N^2}\right)\left(\sup_{k\neq l\in U}n\left|1-\frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\sum_{k\in U}\pi_k\left(\frac{1}{\pi_k}-\frac{N}{n}\right)^2. \qquad (A.5)$$

From the inequality

$$\sum_{k\in U}\pi_k\left(\frac{1}{\pi_k}-\frac{N}{n}\right)^2 \leq \sum_{k\in U}\frac{1}{\pi_k}, \qquad (A.6)$$

we obtain

$$E_p(T_1^2) \leq \left(\frac{1}{N^2}\right)\left(\sup_{k\neq l\in U}n\left|1-\frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\sum_{k\in U}\frac{1}{\pi_k}. \qquad (A.7)$$

By a similar proof, we obtain

$$E_p(T_2^2) \leq \left(\frac{1}{N^2}\right)\left(\sup_{k\neq l\in U}n\left|1-\frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\sum_{k\in U}\frac{1}{\pi_k}, \qquad (A.8)$$

which leads to (2.5).

B Proof of Proposition 2

Lemma B.1. *We have:*

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

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

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

with

$$T_3 \quad = \quad \frac{1}{N^2}\sum_{k\in U}\omega_k^2 v_k^{-2}\pi_k p_k(1-\pi_k p_k) tr(z_k z_k^\top z_k z_k^\top), \tag{B.3}$$

$$T_4 \quad = \quad \frac{1}{N^2}\sum_{k\neq l\in U}\omega_k v_k^{-1}\omega_l v_l^{-1}(\pi_{kl} - \pi_k\pi_l) p_k p_l tr(z_k z_k^\top z_l z_l^\top). \tag{B.4}$$

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

$$T_3 \quad \leq \quad \left(\frac{(C_6)^2 C_2}{n(C_7)^2}\right)\left(\frac{1}{N}\sum_{k\in U}\|z_k\|^4\right), \tag{B.5}$$

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

$$T_4 \quad \leq \quad \frac{1}{N^2}\left(\sup_{k\neq l\in U}\left|1 - \frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\sum_{k\neq l\in U}\omega_k v_k^{-1}\omega_l v_l^{-1}\pi_k\pi_l p_k p_l tr(z_k z_k^\top z_l z_l^\top). \tag{B.6}$$

Since

$$tr(z_k z_k^\top z_l z_l^\top) = (z_k^\top z_l)^2 \leq \|z_k\|^2\|z_l\|^2, \tag{B.7}$$

we obtain

$$\begin{aligned} T_4 &\leq \frac{1}{N^2}\left(\sup_{k\neq l\in U}\left|1 - \frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\left(\sum_{k\in U}\omega_k v_k^{-1}\pi_k p_k\|z_k\|^2\right)^2 \\ &\leq \frac{(C_2)^2(C_6)^2}{n(C_7)^2}\left(\sup_{k\neq l\in U} n\left|1 - \frac{\pi_{kl}}{\pi_k\pi_l}\right|\right)\left(\frac{1}{N}\sum_{k\in U}\|z_k\|^2\right)^2, \end{aligned} \tag{B.8}$$

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

We now consider the proof of Proposition 2. We can write

$$\hat{B}_{ar} - \beta \;=\; T_5 + T_6, \tag{B.9}$$

where

$$T_5 \;=\; \hat{G}_{ar}^{-1} \left\{ \frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k (y_k - z_k^\top \beta) \right\}, \tag{B.10}$$

$$T_6 \;=\; \hat{G}_{ar}^{-1} \left\{ (\hat{G}_r - \hat{G}_{ar}) 1(\hat{G}_{ar} \neq \hat{G}_r) \right\} \beta. \tag{B.11}$$

We first consider the term T_5. We have:

$$\begin{aligned} \|T_5\|^2 \;&\leq\; \|\hat{G}_{ar}^{-1}\|^2 \times \left\| \frac{1}{N} \sum_{k \in S} r_k \omega_k v_k^{-1} z_k (y_k - z_k^\top \beta) \right\|^2 \\ &\leq\; a^{-2} \times \frac{1}{N^2} \sum_{k,l \in S} r_k r_l \omega_k \omega_l v_k^{-1} v_l^{-1} z_k^\top z_l (y_k - z_k^\top \beta)(y_l - z_l^\top \beta), \end{aligned} \tag{B.12}$$

where the second line in (B.12) follows from (3.10). Since the sampling design is non-informative and the response mechanism is unconfounded, we can write

$$E(\|T_5\|^2) \;=\; E_{pq} E_m (\|T_5\|^2), \tag{B.13}$$

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.12), (B.13), and from the assumptions on the imputation model (3.3), we obtain

$$\begin{aligned} E(\|T_5\|^2) \;&\leq\; E_{pq} \left\{ \sigma^2 a^{-2} \times \frac{1}{N^2} \sum_{k \in S} r_k \omega_k^2 v_k^{-1} z_k^\top z_k \right\} \\ &\leq\; \left(\frac{\sigma^2 a^{-2} (C_6)^2 C_2}{n C_7} \right) \left(\frac{1}{N} \sum_{k \in U} \|z_k\|^2 \right) \end{aligned} \tag{B.14}$$

where the second line in (B.14) 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:

$$\begin{aligned} \|T_6\|^2 \;&\leq\; \|\hat{G}_{ar}^{-1}\|^2 \times \left\| (\hat{G}_r - \hat{G}_{ar}) 1(\hat{G}_{ar} \neq \hat{G}_r) \right\|^2 \times \|\beta\|^2 \\ &\leq\; a^{-2} \|\beta\|^2 \times \left\| (\hat{G}_r - \hat{G}_{ar}) 1(\hat{G}_{ar} \neq \hat{G}_r) \right\|^2. \end{aligned} \tag{B.15}$$

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

$$E(\|T_6\|^2) \leq \|\beta\|^2 \times Pr(\hat{G}_{ar} \neq \hat{G}_r). \tag{B.16}$$

We write

$$G = \sum_{j=1}^{p} \eta_j u_j u_j^\top, \tag{B.17}$$

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

$$\begin{aligned} Pr(\hat{G}_{ar} \neq \hat{G}_r) &= Pr(\eta_{pr} \neq a) \\ &\leq Pr\left(|\eta_{pr} - \eta_p| \geq \frac{|\eta_p - a|}{2}\right) \\ &\leq \frac{4}{(\eta_p - a)^2} E(|\eta_{pr} - \eta_p|^2) \qquad \text{(B.18)} \\ &\leq \frac{4}{(\eta_p - a)^2} E\|\hat{G}_r - G\|^2 \qquad \text{(B.19)} \end{aligned}$$

where equation (B.18) follows from the Chebyshev inequality, and equation (B.19) 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.16) and (B.19), and using Lemma B.1, we obtain $E(\|T_6\|^2) = O(n^{-1})$. This completes the proof.

C Proof of Proposition 3

From equation (2.4) in Proposition 1, we obtain under Assumptions (H1), (H2), (H5) and under the model assumptions that

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

It is therefore sufficient to prove that

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

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

$$T_7 = N^{-1} \sum_{k \in S} d_k (1 - r_k) z_k^\top (\hat{B}_{ar} - \beta), \tag{C.3}$$

$$T_8 = \sigma N^{-1} \sum_{k \in S} d_k (1 - r_k) v_k^{1/2} \epsilon_k. \tag{C.4}$$

We have

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

where the second line in (C.5) follows from Assuptions (H1) and (H5). From Proposition 2, 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

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

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

D Proof of Proposition 4

We can write

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

where

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

and $y_k^{**} = z_k^\top \hat{B}_{ar} + \hat{\sigma} v_k^{1/2} \epsilon_k^{**}$ is the imputed value under the balanced random imputation procedure of Chauvet et al. (2011), see equation (4.6).

Since the number of units such that $0 < \psi_{kl}^* < 1$ 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

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

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.