Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16 No. 1, 2018 (New Series), pp 123-144

Joint Imputation Procedures for Categorical Variables

Hélène Chaput[1], Guillaume Chauvet[2], David Haziza[3], Laurianne Salembier[4] and Julie Solard[5]

[1] *Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques, Bureau des professions de santé, France*

[2] *Ensai (Irmar), Campus de Ker Lann, Bruz, France*

[3] *Department of mathematics and statistics, Université de Montréal, Canada*

[4] *Insee, Direction des statistiques démographiques et sociales, division salaires et revenus d'activité*

[5] *Ministère de l'Education Nationale, Direction de l'évaluation, de la prospective et de la performance, France*

Received: November 20, 2017; Revised: April 21, 2018; Accepted: May 8, 2018

Abstract

Marginal imputation, which consists of imputing each item requiring imputation separately, is often used in surveys. This type of imputation procedures leads to asymptotically unbiased estimators of simple parameters such as population totals (or means), but tends to distort relationships between variables. As a result, it generally leads to biased estimators of bivariate parameters such as coefficients of correlation or odd-ratios. Household and social surveys typically collect categorical variables, for which missing values are usually handled by nearest-neighbour imputation or random hot-deck imputation. In this paper, we propose a simple random imputation procedure, closely related to random hot-deck imputation, which succeeds in preserving the relationship between categorical variables. Also, a fully efficient version of the latter procedure is proposed. A limited simulation study compares several estimation procedures in terms of relative bias and relative efficiency.

Key words: balanced random imputation; coefficient of correlation; categorical variable; fully efficient estimator; joint proportion; odd-ratio; random hot-deck imputation.

1 Introduction

Single imputation, which consists of replacing a missing value by an artificial value, is often used in statistical agencies for treating item nonresponse. The main objective of imputation is to reduce the nonresponse bias, which may be appreciable when the respondents and non-respondents differ with respect to the study variables. Achieving an efficient bias reduction relies on the availability of auxiliary information, which is a set of variables observed for all the sample units. Imputation leads to a complete rectangular data file, which is attractive for an analyst since complete data estimation methods may be readily applied to compute point estimates. In some cases, response

Corresponding Author: David Haziza
Email id:

flags, indicating the item specific response statuses for each unit, are provided in the imputed data file. In some situations, however, the flags are not provided by statistical agencies.

In household and social surveys, missing values are often handled through donor imputation procedures such as nearest-neighbour imputation or random hot-deck imputation. In this paper, we focus on survey weighted random hot-deck imputation, whereby a missing value is imputed by the value of a respondent (donor) selected at random from the set of respondents with probabilities proportional to their sampling weights. In practice, survey weighted random hot-deck imputation is generally applied independently within imputation classes, defined on the basis of auxiliary information. The reader is referred to Andridge and Little (2010) for more details on random hot-deck imputation; for multiple imputation methods suitable for categorical variables, see for example White et al. (2010), Van Buuren (2012) and the references therein.

Most often, survey statisticians are interested in estimating simple parameters such as population totals, means and marginal proportions. In this case, marginal imputation, which consists of imputing variables separately, leads to asymptotically unbiased estimators, provided that the assumed imputation model is correctly specified (Haziza, 2009). For example, one may use random hot-deck imputation for each variable requiring imputation. However, this type of method tends to distort the relationships between variables. As a result, estimators of parameters measuring the relationship between variables may be severely biased, especially if the nonresponse rates are appreciable. It is thus desirable to develop imputation strategies which succeed in preserving the relationship between categorical variables. For bivariate parameters involving continuous variables, Shao and Wang (2002) proposed a joint random regression imputation procedure and showed that it leads to asymptotically unbiased estimators of correlation coefficients. Chauvet and Haziza (2012) proposed a fully efficient version of the Shao-Wang procedure in the sense that the imputation variance is eliminated or considerably reduced. A different approach for dealing with bivariate parameters was considered in Skinner and Rao (2002), who proposed to first use marginal imputation to fill in the missing values and then to adjust for the bias at the estimation stage.

In household and social surveys, variables are often categorical so that the methods described above are not directly applicable: rather than dealing with means and correlations, we are interested in marginal and joint proportions. We propose a simple joint random hot-deck imputation procedure that requires the same amount of information that is needed for random hot-deck imputation, and show that it preserves the relationship between categorical variables in the sense that imputed estimators of the joint proportions are approximately unbiased for their population counterparts. Also, a balanced version is proposed, for which the imputation variance is virtually eliminated. The proposed procedure leads to efficient and approximately unbiased estimators of joint proportions while being more efficient than random hot-deck imputation if the interest lies in estimating the marginal proportions.

2 Set-up

Consider a finite population U of size N. Let x denote a categorical study variable with possible characteristics $k = 0, \ldots, K-1$. Similarly, let y denote a categorical study variable with possible

characteristics $l = 0, \ldots, L-1$. We are interested in estimating $p_{k\bullet} = N^{-1} \sum_{i \in U} 1(x_i = k)$, the marginal proportion of units who possess the characteristic k for x; $p_{\bullet l} = N^{-1} \sum_{i \in U} 1(y_i = l)$, the marginal proportion of units who possess the characteristic l for y; and $p_{kl} = N^{-1} \sum_{i \in U} 1(x_i = k)1(y_i = l)$, the joint proportion of units who possess both characteristics k for x and l for y.

A sample s of size n is selected from U according to some sampling design $p(\cdot)$. Let $w_i = 1/\pi_i$ be the sampling weight attached to unit i, where $\pi_i = P(i \in s)$ denotes its first-order inclusion probability in the sample. Complete data estimators of $p_{k\bullet}$, $p_{\bullet l}$ and p_{kl} are the Horvitz-Thompson (1952) estimators

$$\begin{aligned} \hat{p}_{k\bullet} &= N^{-1} \sum_{i \in s} w_i 1(x_i = k), \\ \hat{p}_{\bullet l} &= N^{-1} \sum_{i \in s} w_i 1(y_i = l), \\ \hat{p}_{kl} &= N^{-1} \sum_{i \in s} w_i 1(x_i = k)1(y_i = l). \end{aligned} \tag{2.1}$$

The estimators $\hat{p}_{k\bullet}$, $\hat{p}_{\bullet l}$ and $\hat{p}_{kl}$ are design-unbiased for $p_{k\bullet}$, $p_{\bullet l}$ and p_{kl}, respectively. That is,

$$\begin{aligned} E_p(\hat{p}_{k\bullet}) &= p_{k\bullet}, \\ E_p(\hat{p}_{\bullet l}) &= p_{\bullet l}, \\ E_p(\hat{p}_{kl}) &= p_{kl}, \end{aligned}$$

where $E_p(\cdot)$ denotes the expectation with respect to the sampling design. Alternatively, the denominator $N = \sum_{i \in U} 1$ in (2.1) can be estimated by $\widehat{N} = \sum_{i \in s} w_i$, which leads to the so-called Hàjek estimators of $p_{k\bullet}$, $p_{\bullet l}$ and p_{kl} (Hàjek, 1971). For simplicity, we confine to the case of the Horvitz-Thompson estimators given by (2.1). In practice, both x and y are prone to missing values and require some form of imputation.

In this paper, we assume that the units respond independently of one another. Also, the finite population U is assumed to be partitioned into G imputation classes $U^1, \ldots, U^g, \ldots, U^G$ of size $N^1, \ldots, N^g, \ldots, N^G$, respectively. In class U^g, denote by $s^g = s \cap U^g$ the sample members; s^g_{rr} the set of n^g_{rr} respondents to both items x and y; s^g_{rm} the set of n^g_{rm} respondents to item x only; s^g_{mr} the set of n^g_{mr} respondents to item y only; s^g_{mm} the set of n^g_{mm} non-respondents to both items. Let $\phi^g_{i\circ}$ denote $P(i \in s^g_\circ | i \in s)$ for any response/nonresponse pattern $\circ \in \{rr, rm, mr, mm\}$. We assume that a given pattern occurs with the same probability for any unit $i \in s^g$, so that we simplify the notation as $\phi^g_{i\circ} = \phi^g_\circ$. The data are thus assumed to be Missing Completely At Random (MCAR) within the imputation classes.

In practice, we may ensure that a given pattern occurs with (approximately) the same probability inside an imputation class, by building these imputation classes as follows. We first select the auxiliary variables that are related to the probability of response to x and y. We then fit a polytomous logistic regression model using the selected auxiliary variables as predictors. For sample unit i, we obtain the vector of estimated response probabilities $(\hat{\phi}_{irr}, \hat{\phi}_{irm}, \hat{\phi}_{imr}, \hat{\phi}_{imm})^\top$. Based on these vectors, the sample is then partitioned into homogeneous groups by using a classification

algorithm such as the k-means algorithm. Each of these groups forms one imputation class, and the estimated response probability $\hat{\phi}^g_\circ$ for each pattern $\circ$ inside the imputation class g is simply taken as the frequency of this pattern inside g. This method can be viewed as an extension of the so-called score method (Haziza and Beaumont, 2007) to the case of two study variables.

The population proportions of interest may be rewritten as

$$p_{k\bullet} = N^{-1}\sum_{g=1}^{G} N^g\, p^g_{k\bullet} \quad \text{with} \quad p^g_{k\bullet} = (N^g)^{-1}\sum_{i\in U^g} 1(x_i = k),$$

$$p_{\bullet l} = N^{-1}\sum_{g=1}^{G} N^g\, p^g_{\bullet l} \quad \text{with} \quad p^g_{\bullet l} = (N^g)^{-1}\sum_{i\in U^g} 1(y_i = l),$$

$$p_{kl} = N^{-1}\sum_{g=1}^{G} N^g\, p^g_{kl} \quad \text{with} \quad p^g_{kl} = (N^g)^{-1}\sum_{i\in U^g} 1(x_i = k)1(y_i = l).$$

Similarly, the complete data estimators (2.1) may be rewritten as

$$\hat{p}_{k\bullet} = N^{-1}\sum_{g=1}^{G} \hat{N}^g\, \hat{p}^g_{k\bullet} \quad \text{with} \quad \hat{p}^g_{k\bullet} = (\hat{N}^g)^{-1}\sum_{i\in s^g} w_i 1(x_i = k),$$

$$\hat{p}_{\bullet l} = N^{-1}\sum_{g=1}^{G} \hat{N}^g\, \hat{p}^g_{\bullet l} \quad \text{with} \quad \hat{p}^g_{\bullet l} = (\hat{N}^g)^{-1}\sum_{i\in s^g} w_i 1(y_i = l),$$

$$\hat{p}_{kl} = N^{-1}\sum_{g=1}^{G} \hat{N}^g\, \hat{p}^g_{kl} \quad \text{with} \quad \hat{p}^g_{kl} = (\hat{N}^g)^{-1}\sum_{i\in s^g} w_i 1(x_i = k)1(y_i = l),$$

where $\hat{N}^g = \sum_{i\in s^g} w_i$ is an estimator of the g-th class size, N_g.

Let x^*_i and y^*_i be the imputed values used to replace the missing x_i and y_i. Imputed estimators of $p_{k\bullet}$, $p_{\bullet l}$ and p_{kl} are respectively

$$\begin{aligned}
\hat{p}_{k\bullet,I} &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{r\bullet}} w_i 1(x_i = k) + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{m\bullet}} w_i 1(x^*_i = k),\\
\hat{p}_{\bullet l,I} &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{\bullet r}} w_i 1(y_i = l) + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{\bullet m}} w_i 1(y^*_i = l), \qquad (2.2)\\
\hat{p}_{kl,I} &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{rr}} w_i 1(x_i = k)1(y_i = l) + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{rm}} w_i 1(x_i = k)1(y^*_i = l)\\
&+ N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mr}} w_i 1(x^*_i = k)1(y_i = l) + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mm}} w_i 1(x^*_i = k)1(y^*_i = l),
\end{aligned}$$

where $s^g_{r\bullet} = s^g_{rr} \cup s^g_{rm}$ denotes the set of respondents to item x in class g; $s^g_{m\bullet} = s^g_{mr} \cup s^g_{mm}$ denotes the set of non-respondents to item x in class g, and $s^g_{\bullet r}$ and $s^g_{\bullet m}$ corresponding to item y are

similarly defined. Once the data have been imputed, the computation of (2.2) does not require the response flags to be available in the imputed data file. Complete data estimation procedures may thus be readily applied by secondary analysts for point estimation, which is an important practical aspect.

In order to study the properties of an imputed estimator $\hat{p}_{\diamond,I}$ of a proportion $p_\diamond$, we express its total error as

$$\hat{p}_{\diamond,I} - p_\diamond \;=\; (\hat{p}_\diamond - p_\diamond) + (\tilde{p}_{\diamond,I} - \hat{p}_\diamond) + (\hat{p}_{\diamond,I} - \tilde{p}_{\diamond,I})\,, \tag{2.3}$$

where $\tilde{p}_{\diamond,I} \equiv E_I\left(\hat{p}_{\diamond,I}\right)$, for $\diamond \in \{k\bullet, \bullet l, kl\}$, and $E_I(\cdot)$ denotes the expectation with respect to the imputation mechanism, conditionally on the sample s and on the sets of respondents to items x and y. In other words, $E_I(\cdot)$ denotes the expectation with respect to the random selection of donors in the case of a random imputation method. The first term on the right hand side of (2.3) represents the sampling error, whereas the second and the third terms represent the non-response error and the imputation error. The imputation error occurs solely from the random imputation mechanism. We seek an imputation procedure under which the non-response bias

$$B_{pqI}(\hat{p}_{\diamond,I}) \;\equiv\; E_p E_q E_I\left(\hat{p}_{\diamond,I} - \hat{p}_\diamond\right) = E_p E_q\left(\tilde{p}_{\diamond,I} - \hat{p}_\diamond\right)$$

is approximately equal to 0, where $E_q(\cdot)$ denotes the expectation with respect to the assumed non-response model, conditionally on the sample s.

We focus on survey weighted random hot-deck imputation, which consists of selecting a donor at random from the set of respondents with probability proportional to its sampling weight, and then using the donor's item value(s) to "fill in" for the missing value of a non-respondent. Marginal random hot-deck imputation, which consists of imputing x and y separately, tends to attenuate the relationship between items being imputed. As a result, this method introduces a bias in the estimation of p_{kl} that may be severe if the non-response rate is appreciable. In practice, it is customary to use a slightly different version of random hot-deck imputation that consists of using a common donor when both x and y are missing. For any class U^g, we proceed as follows:

(i) for $i \in s^g_{mr}$, missing x_i is imputed by $x^*_i = k$ with probability

$$\hat{p}^g_{k\bullet,ac} \equiv (\hat{N}^g_{r\bullet})^{-1} \sum_{i \in s^g_{r\bullet}} w_i 1(x_i = k) \tag{2.4}$$

estimated from the available cases (AC) in class g for item x, and $\hat{N}^g_{r\bullet} = \sum_{i \in s^g_{r\bullet}} w_i$;

(ii) for $i \in s^g_{rm}$, missing y_i is imputed by means of an analogous procedure;

(iii) for $i \in s^g_{mm}$, missing (x_i, y_i) is imputed by $(x^*_i, y^*_i) = (k, l)$ with probability

$$\hat{p}^g_{kl,cc} \equiv (\hat{N}^g_{rr})^{-1} \sum_{i \in s^g_{rr}} w_i 1(x_i = k) 1(y_i = l) \tag{2.5}$$

estimated from the complete cases (CC) in class g to items x and y, with $\hat{N}^g_{rr} = \sum_{i \in s^g_{rr}} w_i$.

When one variable only is missing, random hot-deck imputation estimates its distribution separately from available cases for this variable. When both variables are missing, their distribution is estimated jointly from complete cases for both variables. Random hot-deck imputation succeeds in preserving the marginal distributions of x and y. Therefore, $B_{pqI}(\hat{p}_{k\bullet,I}) \simeq 0$ and $B_{pqI}(\hat{p}_{\bullet l,I}) \simeq 0$ for any characteristics k and l. Although this imputation procedure generates less bias than marginal random hot-deck imputation, there generally remains some bias when estimating the joint proportions, since

$$B_{pqI}(\hat{p}_{kl,I}) \simeq -N^{-1} \sum_{g=1}^{G} N^g (\phi^g_{rm} + \phi^g_{mr})(p^g_{kl} - p^g_{k\bullet} p^g_{\bullet l}). \tag{2.6}$$

The proof of (2.6) is given in Appendix A. The asymptotic bias vanishes if $\phi^g_{rm} = \phi^g_{mr} = 0$ for any g, which means that items x are y may not be missing separately, or if both x and y are unrelated within imputation classes, in which case $p^g_{kl} = p^g_{k\bullet} p^g_{\bullet l}$.

3 Proposed Imputation Procedures

To account for the existing relationship between variables, we propose two imputation procedures, where the distribution of x is estimated conditionally on y if x only is missing, and where the distribution of y is estimated conditionally on x if y only is missing. For any unit $i \in U^g$, we note

$$\hat{p}^g_{k|l,cc} = \frac{\sum_{i \in s^g_{rr}} w_i 1(x_i = k) 1(y_i = l)}{\sum_{i \in s^g_{rr}} w_i 1(y_i = l)}$$

the estimated probability that $x_i = k$ when $y_i = l$, and

$$\hat{p}^g_{l|k,cc} = \frac{\sum_{i \in s^g_{rr}} w_i 1(x_i = k) 1(y_i = l)}{\sum_{i \in s^g_{rr}} w_i 1(x_i = k)}$$

the estimated probability that $y_i = l$ when $x_i = k$.

As pointed out by Chauvet et al. (2011) and Chauvet and Haziza (2012), imputing missing values may be performed by sampling within populations of cells, separately for each of the sub-samples s^g_{mr}, s^g_{rm} and s^g_{mm}.

(i) To handle units in s^g_{mr}, we create a population of cells U^{g*}_{mr} of size $n^g_{mr} \times K$. Each cell (i,k) is assigned the probability of selection $\hat{p}^g_{k|y_i,cc}$. A random sample s^{g*}_{mr} of size n^g_{mr} is selected from U^{g*}_{mr}, and missing x_i is imputed by $x^*_i = k$ if the cell (i,k) is selected.

(ii) To handle units in s^g_{rm}, we create a population of cells U^{g*}_{rm} of size $n^g_{rm} \times L$. Each cell (i,l) is assigned the probability of selection $\hat{p}^g_{l|x_i,cc}$. A random sample s^{g*}_{rm} of size n^g_{rm} is selected from U^{g*}_{rm}, and missing y_i is imputed by $y^*_i = l$ if the cell (i,l) is selected.

(iii) To handle units in s^g_{mm}, we create a population of cells U^{g*}_{mm} of size $n^g_{mm} \times (KL)$. Each cell (i,q') is assigned the probability of selection $\hat{p}^g_{k_{q'} l_{q'},cc}$. A random sample s^{g*}_{mm} of size n^g_{mm} is selected from U^{g*}_{mm}, and missing (x_i, y_i) is imputed by $(x^*_i, y^*_i) = (k_q, l_q)$ if the cell (i,q) is selected.

In the populations U_{mr}^{g*}, U_{rm}^{g*} and U_{mm}^{g*}, each row stands for a non-respondent, and each column for a possible imputed value. We impose that

C1: the samples s_{mr}^{g*}, s_{rm}^{g*} and s_{mm}^{g*} are drawn so that exactly one cell per row is selected.

The constraint C1 is required since exactly one imputed value must be selected for each non-respondent. Imposing only the constraint C1 results in the *joint random hot-deck imputation* procedure which may be alternatively described as follows:

(i) for $i \in s_{mr}^{g}$, missing x_i is imputed by $x_i^* = k$ with probability $\hat{p}_{k|y_i,cc}^{g}$,

(ii) for $i \in s_{rm}^{g}$, missing y_i is imputed by $y_i^* = l$ with probability $\hat{p}_{l|x_i,cc}^{g}$,

(iii) for $i \in s_{mm}^{g}$, missing (x_i, y_i) is imputed by $(x_i^*, y_i^*) = (k, l)$ with probability $\hat{p}_{kl,cc}^{g}$.

It is shown in Appendix B that $B_{pqI}(\hat{p}_{\diamond,I}) \simeq 0$ under this imputation procedure, for $\diamond \in \{k\bullet, \bullet l, kl\}$ and any characteristics k and l. Guidelines are given in Appendix C to extend the joint random hot-deck imputation procedure to the case of more than two missing items. A drawback of the proposed procedure is that it suffers from an additional variability, called the imputation variance, due to the random selection of donors. To eliminate the imputation variance, we further impose that

C2: the samples s_{mr}^{g*}, s_{rm}^{g*} and s_{mm}^{g*} are drawn so that the following balancing equations are satisfied:

$$\sum_{(i,k)\in s_{mr}^{g*}} \left(\hat{p}_{k|y_i,cc}^{g}\right)^{-1} \mathbf{t}_{ik} \;=\; \sum_{(i,k)\in U_{mr}^{g*}} \mathbf{t}_{ik} \tag{3.1}$$

with $\mathbf{t}_{ik} = \{(\mathbf{t}_{ik})_1, \ldots, (\mathbf{t}_{ik})_{KL}\}^\top$ and $(\mathbf{t}_{ik})_q = w_i\, \hat{p}_{k|y_i,cc}^{g}\, 1(k = k_q)\, 1(y_i = l_q)$ for any $q = 1, \ldots, KL$, where k_q and l_q are the two integers such that $q = k_q \times L + (l_q + 1)$;

$$\sum_{(i,l)\in s_{rm}^{g*}} \left(\hat{p}_{l|x_i,cc}^{g}\right)^{-1} \mathbf{t}_{il} \;=\; \sum_{(i,l)\in U_{rm}^{g*}} \mathbf{t}_{il}, \tag{3.2}$$

with $\mathbf{t}_{il} = \{(\mathbf{t}_{il})_1, \ldots, (\mathbf{t}_{il})_{KL}\}^\top$ and $(\mathbf{t}_{il})_q = w_i\, \hat{p}_{l|x_i,cc}^{g}\, 1(x_i = k_q)\, 1(l = l_q)$;

$$\sum_{(i,q)\in s_{mm}^{g*}} \left(\hat{p}_{k_q l_q,cc}^{g}\right)^{-1} \mathbf{t}_{iq} \;=\; \sum_{(i,q)\in U_{mm}^{g*}} \mathbf{t}_{iq}. \tag{3.3}$$

with $\mathbf{t}_{iq'} = \{(\mathbf{t}_{iq'})_1, \ldots, (\mathbf{t}_{iq'})_{KL}\}^\top$ and $(\mathbf{t}_{iq'})_q = w_i\, \hat{p}_{k_{q'} l_{q'},cc}^{g}\, 1(k_{q'} = k_q)\, 1(l_{q'} = l_q)$.

If the constraint C2 is exactly satisfied, we prove in Appendix D that $\hat{p}_{\diamond,I} - \tilde{p}_{\diamond,I} = 0$ for $\diamond \in \{k\bullet, \bullet l, kl\}$ and any characteristics k and l. As a result, the imputation error in (2.3) is equal to zero and the imputation variance vanishes. If both constraints C1 and C2 are imposed in the selection of cells, we obtain the *balanced joint random hot-deck imputation* procedure. The constraints C1 and C2 may be satisfied by selecting the samples s_{mr}^{g*}, s_{rm}^{g*} and s_{mm}^{g*} by means of the cube method originally developed in the context of balanced sampling; see Deville and Tillé (2004) and Chauvet et al. (2011). The extension of the above procedure to the case of three categorical procedures is presented in Appendix C.

4 Alternative Estimators

We now present some alternative estimation procedures for $p_{k\bullet}$, $p_{\bullet l}$ and p_{kl}. In Section 7, these procedures are compared empirically to the methods described in Sections 2 and 3 in terms of bias and relative efficiency. We start by the complete case (CC) estimators

$$\begin{aligned}
\hat{p}_{k\bullet,cc} &= \hat{N}_{rr}^{-1} \sum_{g=1}^{G} \hat{N}_{rr}^{g}\, \hat{p}_{k\bullet,cc}^{g} \quad \text{with} \quad \hat{p}_{k\bullet,cc}^{g} = (\hat{N}_{rr}^{g})^{-1} \sum_{i \in s_{rr}^{g}} w_i 1(x_i = k), \\
\hat{p}_{\bullet l,cc} &= \hat{N}_{rr}^{-1} \sum_{g=1}^{G} \hat{N}_{rr}^{g}\, \hat{p}_{\bullet l,cc}^{g} \quad \text{with} \quad \hat{p}_{\bullet l,cc}^{g} = (\hat{N}_{rr}^{g})^{-1} \sum_{i \in s_{rr}^{g}} w_i 1(y_i = l), \\
\hat{p}_{kl,cc} &= \hat{N}_{rr}^{-1} \sum_{g=1}^{G} \hat{N}_{rr}^{g}\, \hat{p}_{kl,cc}^{g} \quad \text{with} \quad \hat{p}_{kl,cc}^{g} = (\hat{N}_{rr}^{g})^{-1} \sum_{i \in s_{rr}^{g}} w_i 1(x_i = k) 1(y_i = l),
\end{aligned} \tag{4.1}$$

which are based on the responding units to both x and y, where $\hat{N}_{rr} = \sum_{g=1}^{G} \hat{N}_{rr}^{g}$. The bias of CC estimators can be approximated by

$$\begin{aligned}
B_{pq}(\hat{p}_{k\bullet,cc}) &\simeq \frac{\sum_{g=1}^{G} N_g \{\phi_{rr}^{g} - \bar{\phi}_{rr}\}\{p_{k\bullet}^{g} - p_{k\bullet}\}}{\sum_{g=1}^{G} N_g \phi_{rr}^{g}}, \\
B_{pq}(\hat{p}_{\bullet l,cc}) &\simeq \frac{\sum_{g=1}^{G} N_g \{\phi_{rr}^{g} - \bar{\phi}_{rr}\}\{p_{\bullet l}^{g} - p_{\bullet l}\}}{\sum_{g=1}^{G} N_g \phi_{rr}^{g}}, \\
B_{pq}(\hat{p}_{kl,cc}) &\simeq \frac{\sum_{g=1}^{G} N_g \{\phi_{rr}^{g} - \bar{\phi}_{rr}\}\{p_{kl}^{g} - p_{kl}\}}{\sum_{g=1}^{G} N_g \phi_{rr}^{g}},
\end{aligned} \tag{4.2}$$

where $B_{pq}(\cdot)$ denotes the bias under both the sampling design and the non-response model, and $\bar{\phi}_{rr} = N^{-1} \sum_{g=1}^{G} N_g \phi_{rr}^{g}$. From (4.2), the CC estimators are biased if there is an association between the probability of responding to both variables and the proportion we wish to estimate.

The bias of the CC estimators can be removed by accounting for class information. This leads to the adjusted complete case (ACC) estimators

$$\begin{aligned}
\hat{p}_{k\bullet,acc} &= N^{-1} \sum_{g=1}^{G} \hat{N}^{g}\, \hat{p}_{k\bullet,cc}^{g}, \\
\hat{p}_{\bullet l,acc} &= N^{-1} \sum_{g=1}^{G} \hat{N}^{g}\, \hat{p}_{\bullet l,cc}^{g}, \\
\hat{p}_{kl,acc} &= N^{-1} \sum_{g=1}^{G} \hat{N}^{g}\, \hat{p}_{kl,cc}^{g}.
\end{aligned} \tag{4.3}$$

It can be shown that $B_{pq}(\hat{p}_{\diamond,acc}) \simeq 0$ for any $\diamond \in \{k\bullet, \bullet l, kl\}$. The ACC estimators may be viewed as propensity score adjusted estimators, where the response probability of a unit in a given imputation class is estimated by the response rate to both items within the same class. However,

implementing ACC estimators in order to obtain a complete imputed data file will necessarily lead to "impossible values". For example, in the case of a binary variable (with possible values 0 and 1), the imputed values will never be equal to either 0 or 1 but will lie in the interval $(0, 1)$, which is a drawback from a micro-data point of view. In contrast, the imputation procedures described in Sections 3 and 4 use the values of donors to replace the missing values, which eliminates the problem of impossible values.

Another set of estimators are based on available cases, which leads to the available case (AC) estimators

$$\begin{aligned}
\hat{p}_{k\bullet,ac} &= \hat{N}_{r\bullet}^{-1} \sum_{g=1}^{G} \hat{N}_{r\bullet}^{g}\, \hat{p}_{k\bullet,ac}^{g} \text{ with } \hat{p}_{k\bullet,ac}^{g} = (\hat{N}_{r\bullet}^{g})^{-1} \sum_{i \in s_{r\bullet}^{g}} w_i 1(x_i = k), \\
\hat{p}_{\bullet l,ac} &= \hat{N}_{\bullet r}^{-1} \sum_{g=1}^{G} \hat{N}_{\bullet r}^{g}\, \hat{p}_{\bullet l,ac}^{g} \text{ with } \hat{p}_{\bullet l,ac}^{g} = (\hat{N}_{\bullet r}^{g})^{-1} \sum_{i \in s_{\bullet r}^{g}} w_i 1(y_i = l), \\
\hat{p}_{kl,ac} &= \hat{p}_{kl,cc},
\end{aligned} \tag{4.4}$$

where $\hat{N}_{r\bullet} = \sum_{g=1}^{G} \hat{N}_{r\bullet}^{g}$, and $\hat{N}_{\bullet r}$ is defined similarly. The bias of AC estimators can be approximated by

$$\begin{aligned}
B_{pq}(\hat{p}_{k\bullet,ac}) &\simeq \frac{\sum_{g=1}^{G} N_g \{\phi_{r\bullet}^{g} - \bar{\phi}_{r\bullet}\}\{p_{k\bullet}^{g} - p_{k\bullet}\}}{\sum_{g=1}^{G} N_g \phi_{r\bullet}^{g}}, \\
B_{pq}(\hat{p}_{\bullet l,ac}) &\simeq \frac{\sum_{g=1}^{G} N_g \{\phi_{\bullet r}^{g} - \bar{\phi}_{\bullet r}\}\{p_{\bullet l}^{g} - p_{\bullet l}\}}{\sum_{g=1}^{G} N_g \phi_{\bullet r}^{g}}, \\
B_{pq}(\hat{p}_{kl,ac}) &\simeq \frac{\sum_{g=1}^{G} N_g \{\phi_{rr}^{g} - \bar{\phi}_{rr}\}\{p_{kl}^{g} - p_{kl}\}}{\sum_{g=1}^{G} N_g \phi_{rr}^{g}},
\end{aligned} \tag{4.5}$$

where $\phi_{r\bullet}^{g} = \phi_{rr}^{g} + \phi_{rm}^{g}$ and $\bar{\phi}_{r\bullet} = N^{-1} \sum_{g=1}^{G} N_g \phi_{r\bullet}^{g}$; $\phi_{\bullet r}^{g}$ and $\bar{\phi}_{\bullet r}$ are defined similarly. An AC estimator is thus biased if there exists an association between the probability of responding to the required variables and the proportion we wish to estimate.

The bias can be removed by accounting for class information, which leads to the adjusted available case (AAC) estimators

$$\begin{aligned}
\hat{p}_{k\bullet,aac} &= N^{-1} \sum_{g=1}^{G} \hat{N}^{g}\, \hat{p}_{k\bullet,ac}^{g}, \\
\hat{p}_{\bullet l,aac} &= N^{-1} \sum_{g=1}^{G} \hat{N}^{g}\, \hat{p}_{\bullet l,ac}^{g}, \\
\hat{p}_{kl,aac} &= N^{-1} \sum_{g=1}^{G} \hat{N}^{g}\, \hat{p}_{kl,ac}^{g}.
\end{aligned} \tag{4.6}$$

It can be shown that $B_{pq}(\hat{p}_{\diamond,aac}) \simeq 0$ for any $\diamond \in \{k\bullet, \bullet l, kl\}$. As for the ACC estimators, the AAC estimators can be viewed as propensity score adjusted estimators, where the response probability of a unit within an imputation class is estimated by the response rate based on available respondents within the same class. Also, as for the ACC estimators, the AAC estimators will necessarily lead to impossible values.

5 Variance Estimation under the Balanced Procedure

In this section, we turn our attention to estimating the variance of the imputed estimators under the proposed balanced imputation procedure described in Section 3. It is well known that treating the imputed values as if they were observed leads to serious underestimation of the variance of imputed estimators if the proportion of missing data is appreciable and to poor confidence intervals. Several variance estimation methods accounting for nonresponse and imputation have been proposed in the literature; see Haziza (2009) for a review. In this paper, we focus on the bootstrap method, which was studied by Shao and Sitter (1996). The rationale behind the Shao-Sitter method is to select, using any complete data bootstrap method, a bootstrap sample consisting of original or rescaled imputed data and their corresponding original response statuses. The bootstrap data with a missing status are then reimputed using the same imputation method that was used in the original sample. The proposed balanced joint random hot-deck imputation procedure entails the application of the procedure within each bootstrap sample, which may be highly computer intensive. A simplified bootstrap method can be used by noting that the imputation variance is virtually eliminated under the proposed balanced imputation procedure. It consists of reimputing the deterministic version of the balanced joint random hot-deck imputation procedure within each bootstrap sample, which is equivalent to re-calculating $\tilde{p}_{\diamond,I} \equiv E_I\left(\hat{p}_{\diamond,I}\right)$ within each bootstrap sample, $\diamond \in \{k\bullet, \bullet l, kl\}$. After some relatively straightforward algebra, we obtain

$$\begin{aligned}
\tilde{p}_{k\bullet,I} &\simeq N^{-1}\sum_{g=1}^{G}\left[\hat{N}^g_{r\bullet}\hat{p}^g_{k\bullet,ac} + \hat{N}^g_{mr}\hat{p}^g_{k\bullet,mr} + \hat{N}^g_{mm}\hat{p}^g_{k\bullet,cc}\right], \\
\tilde{p}_{\bullet l,I} &\simeq N^{-1}\sum_{g=1}^{G}\left[\hat{N}^g_{\bullet r}\hat{p}^g_{\bullet l,ac} + \hat{N}^g_{rm}\hat{p}^g_{\bullet l,rm} + \hat{N}^g_{mm}\hat{p}^g_{\bullet l,cc}\right], \qquad (5.1) \\
\tilde{p}_{kl,I} &\simeq N^{-1}\sum_{g=1}^{G}\left[(\hat{N}^g_{rr} + \hat{N}^g_{mm})\hat{p}^g_{kl,cc} + \hat{N}^g_{mr}\hat{p}^g_{kl,mr} + \hat{N}^g_{rm}\hat{p}^g_{kl,rm}\right],
\end{aligned}$$

where $\hat{p}^g_{k\bullet,ac}$ and $\hat{p}^g_{\bullet l,ac}$ are given in (4.4), $\hat{p}^g_{k\bullet,cc}$, $\hat{p}^g_{\bullet l,cc}$ and $\hat{p}^g_{kl,cc}$ are given in (4.1) and

$$\begin{aligned}
\hat{p}^g_{k\bullet,mr} &= \frac{\sum_{i\in s^g_{mr}} w_i \sum_{l=1}^{L} 1(y_i = l)\hat{p}^g_{k|l,cc}}{\sum_{i\in s^g_{mr}} w_i}, \\
\hat{p}^g_{\bullet l,rm} &= \frac{\sum_{i\in s^g_{rm}} w_i \sum_{k=1}^{K} 1(x_i = k)\hat{p}^g_{l|k,cc}}{\sum_{i\in s^g_{rm}} w_i}, \\
\hat{p}^g_{kl,mr} &= \frac{\sum_{i\in s^g_{mr}} w_i 1(y_i = l)\hat{p}^g_{k|l,cc}}{\sum_{i\in s^g_{mr}} w_i}, \\
\hat{p}^g_{kl,rm} &= \frac{\sum_{i\in s^g_{rm}} w_i 1(x_i = k)\hat{p}^g_{l|k,cc}}{\sum_{i\in s^g_{rm}} w_i}.
\end{aligned}$$

As an illustration, we use the bootstrap weight method of Rao, Wu and Yue (1992) in the special case of simple random sampling without replacement. The extension to stratified simple random sampling without replacement is straightforward. The bootstrap weight procedure proceeds as follows:

(1) Let n' be the bootstrap sample size, which may be different from n.

(2) Draw a simple random sample *with* replacement s^* of size n' from s. Let m_i^* be the number of times unit i is selected in s^*. We have $n' = \sum_{i \in s} m_i^*$. For unit $i \in s$, define the bootstrap weight as

$$w_i^* = w_i \left\{ 1 + \sqrt{\lambda} \left(\frac{n m_i^*}{n'} - 1 \right) \right\} \quad \text{with} \quad \lambda = \frac{n' \left(1 - \frac{n}{N}\right)}{n - 1}.$$

Compute $\tilde{p}_{\diamond,I}^*$ from (5.1) by replacing w_i with w_i^*.

(3) Repeat Step 2 a large number of times, C, to get with-replacement samples $s^{*(1)}, \ldots, s^{*(C)}$. For each sample $s^{*(c)}$, $c = 1, \ldots, C$, compute $\tilde{p}_{\diamond,I}^{*(c)}$ like in Step 2.

(4) Estimate the variance of $\tilde{p}_{\diamond,I}$ by

$$\hat{V}_{1C} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\tilde{p}_{\diamond,I}^{*(c)} - \frac{1}{C} \sum_{d=1}^{C} \tilde{p}_{\diamond,I}^{*(d)} \right)^2. \tag{5.2}$$

The reader is referred to Chauvet (2007, 2015) for a review of bootstrap methods in survey sampling, and to Antal and Tillé (2011) and Beaumont and Patak (2012) for bootstrap weight methods in the context of unequal probability sampling designs. If the sampling fraction n/N is negligible, the bootstrap variance estimators (5.2) are consistent for the true variance; see Haziza (2009) and Mashreghi et al. (2014) for a discussion on the consistency of the method of Shao and Sitter (1996). Variance estimation for non-negligible sampling fractions in the context of bivariate parameters requires further investigations.

6 Simulation Study

We conducted two simulation studies to test the performance of the point and variance estimation procedures described in Sections 2-5. In the first study, we compared the performance of several point estimation procedures in terms of relative bias and relative efficiency. In the second, we tested the performance of the bootstrap variance estimator described in Section 5.

6.1 Performance of the Point Estimators

We generated a finite population of size $N = 20,000$ consisting of two binary variables x and y so that $k \in \{0, 1\}$ and $l \in \{0, 1\}$. The population consisted of five classes, each of size $4,000$. We were interested in estimating the marginal first moments $p_{1\bullet}$ and $p_{\bullet 1}$, the joint proportion p_{11} as well as the population odd-ratio

$$\mathrm{OR} = \frac{\mathrm{p}_{11}\, \mathrm{p}_{00}}{\mathrm{p}_{10}\, \mathrm{p}_{01}}. \tag{6.1}$$

From the population, we selected $B = 10,000$ samples of size $n = 2,000$ according to simple random sampling without replacement. In each selected sample, non-response to x and y was generated according to a non-response mechanism described in Table 1, along with the population characteristics. The characteristics of the population were chosen so as to obtain a positive association between ϕ^g_{rr} and $p^g_{1\bullet}$, between ϕ^g_{rr} and $p^g_{\bullet 1}$, and between ϕ^g_{rr} and p^g_{11}. The CC estimators are therefore expected to be positively biased; see (4.2). Also, the characteristics of the population were chosen so as to obtain a positive association between $\phi^g_{r\bullet}$ and $p^g_{1\bullet}$, and between $\phi^g_{\bullet r}$ and $p^g_{\bullet 1}$. The AC estimators are therefore expected to be positively biased; see (4.5).

Table 1: Characteristics of the population and mechanism used to generate nonresponse

Class	$p_{1\bullet}$	$p_{\bullet 1}$	p_{11}	OR	ϕ_{rr}	ϕ_{rm}	ϕ_{mr}	ϕ_{mm}
1	0.50	0.50	0.20	0.44	0.10	0.20	0.20	0.50
2	0.55	0.55	0.30	0.96	0.20	0.20	0.20	0.40
3	0.60	0.60	0.40	2.00	0.30	0.25	0.25	0.20
4	0.65	0.65	0.50	4.44	0.40	0.20	0.20	0.20
5	0.70	0.70	0.60	12.00	0.50	0.20	0.20	0.10

In each sample, we computed seven estimators for each of the parameters of interest $p_{k\bullet}$, $p_{\bullet l}$, p_{11} and OR: (i) the CC estimators given in (4.1); (ii) the ACC estimators given in (4.3); (iii) the AC estimators given in (4.4); (iv) the AAC estimators given in (4.6); (v) the imputed estimators given by (2.2) based on the random hot-deck imputation (RHDI) procedure described in Section 2; (vi) the imputed estimators given by (2.2) based on the joint random hot-deck imputation (JHDI) procedure described in Section 3; (vii) the imputed estimators given by (2.2) based on the balanced joint random hot-deck imputation (BJRHDI) procedure described in Section 3. In each case, an estimator $\widehat{\mathrm{OR}}_I$ of the OR was obtained by replacing each unknown parameter in (6.1) by its corresponding imputed estimator. Using a personal laptop and the IML procedure of the SAS software, it took an average time of 0.13 seconds to select a sample and to compute the associated estimators.

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

$$RB(\hat{\theta}) = \frac{E_{MC}(\hat{\theta}) - \theta}{\theta} \times 100, \tag{6.2}$$

where $E_{MC}(\hat{\theta}) = B^{-1} \sum_{b=1}^{B} \hat{\theta}^{(b)}$ and $\hat{\theta}^{(b)}$ denotes the estimator $\hat{\theta}$ in the b-th sample, $b = 1, \ldots, 10\,000$. When the true value of the parameter θ is close to zero, the relative bias may not be an appropriate measure. This is not problematic in our simulation set-up as the values of $p_{1\bullet}$, $p_{\bullet 1}$, p_{11} and OR were bounded away from 0 (see Table 1). As a measure of Relative Efficiency (RE), we used

$$RE = \frac{MSE_{MC}(\hat{\theta}^{(aac)})}{MSE_{MC}(\hat{\theta}^{(\cdot)})} \times 100, \tag{6.3}$$

where $MSE_{MC}(\hat{\theta})$ is the Monte Carlo mean square error of $\hat{\theta}$ and $\hat{\theta}^{aac}$ denote the adjusted available-case estimator.

Table 2: Monte-Carlo percent relative bias and relative efficiency (between brackets) of several estimators

Estimator	$p_{1\bullet}$	$p_{\bullet 1}$	p_{11}	OR
CC	5.6 (15)	5.5 (17)	16.7 (10)	71.2 (28)
ACC	0.0 (46)	0.0 (44)	0.0 (100)	35.6 (100)
AC	3.3 (41)	3.3 (42)	16.7 (10)	71.2 (28)
AAC	0.0 (100)	0.0 (100)	0.0 (100)	35.6 (100)
RHDI	0.0 (68)	0.0 (68)	-3.7 (89)	-21.8 (278)
JHDI	0.0 (60)	0.0 (59)	0.0 (115)	2.5 (329)
BJRHDI	0.0 (70)	0.0 (67)	0.0 (131)	2.3 (377)

Table 2 shows the Monte Carlo percent Relative Bias (RB) and percent Relative Efficiency (RE) of the seven estimators of $p_{1\bullet}$, $p_{\bullet 1}$, p_{11} and OR. The CC estimators and the AC estimators showed positive bias for $p_{1\bullet}$, $p_{\bullet 1}$ and p_{11}, as expected. As a result, the corresponding estimators of OR were strongly biased with a value of RB equal to 71.2%. The ACC estimator and the AAC estimator, which account for class information, showed virtually no bias for $p_{1\bullet}$, $p_{\bullet 1}$ and p_{11}, but were significantly biased for OR with a value of RB equal to 35.6%. Turning to the imputed estimators, we note that the imputed estimators of the marginal proportions showed no bias, as expected. However, under RHDI, both the imputed estimator of p_{11} and the estimator of OR were biased with values of RB equal to -3.7% and -21.6%, respectively. Also, the biases were negative clearly illustrating the problem of attenuation of relationships. On the other hand, both JHDI and BJRHDI led to negligible bias, showing that both procedures succeeded in preserving the relationship between variables.

We now turn to the relative efficiency. We first consider the marginal first moments. We note that the CC and ACC estimators were inefficient, which can be explained by the fact that they tend to discard a lot of information. The imputed estimators under both RHDI and JHDI were less efficient than the corresponding AAC estimator with values of RE ranging from 59% to 68%. The efficiency loss arises from the random selection of donors in the random hot-deck imputation procedures. The imputed estimators under BJRHDI were more efficient than the corresponding estimators obtained under RHDI and JHDI, illustrating the reduction of the imputation variance. In regards to the joint proportion p_{11}, the imputed estimator under RHDI was less efficient than the AAC estimators, while the imputed estimators under both JHDI and BJRHDI were more efficient. The imputed estimator of OR under all three imputation methods was considerably more efficient than the AAC estimator.

6.2 Performance of the Variance Estimators

We conducted a second simulation study on the same population in order to assess the performance of the bootstrap procedure described in Section 5. We were interested in estimating the variance of the marginal first moments $p_{1\bullet}$ and $p_{\bullet 1}$, the joint proportion p_{11} as well as the population odd-ratio OR.

From each population, we selected $B = 10,000$ samples of size $n = 1,000$ according to simple random sampling without replacement. In each selected sample, non-response to x and y was generated according to the non-response mechanism described in Table 1. We were interested in estimating the variance of the imputed estimators of $p_{1\bullet}$, $p_{\bullet 1}$, p_{11} and OR under the proposed balanced joint random hot-deck imputation procedure. In each sample (containing respondents and nonrespondents), we selected $C = 2,000$ bootstrap samples according to the bootstrap weight procedure of Section 5. To measure the bias of the Bootstrap variance estimator, we used the Monte Carlo percent relative bias given by (6.2). The true variance was replaced by a Monte Carlo approximation, obtained through an independent run of $50,000$ simulations. Also, we computed confidence intervals by means of the percentile method. For example, in the case of $\widehat{\mathrm{OR}}_I$, we computed the C bootstrap versions of the odd-ratio, $\tilde{\mathrm{OR}}_I^{*(c)}, c = 1, \ldots, C$. An $(1-2\alpha)$ confidence interval is then given by $\left[\tilde{\mathrm{OR}}_I^{*(L)}, \tilde{\mathrm{OR}}_I^{*(U)}\right]$ with $L = \alpha\, C$ and $U = (1-\alpha)\, C$. We computed the coverage error rates of the percentile bootstrap confidence intervals, with nominal error rates of 2.5% and 5% in each tail. Using a personal laptop and the IML procedure of the SAS software, it took an average time of 22 seconds to select a sample along with $C = 2,000$ bootstrap resamples.

Table 3 shows the Monte Carlo percent relative bias (RB) of the Bootstrap variance estimator and the error rates. The Bootstrap variance estimator performed well for $\hat{p}_{1\bullet,I}$, $\hat{p}_{\bullet 1,I}$ and $\hat{p}_{11,I}$, with an absolute relative bias less than 5%. The Bootstrap variance estimator was positively biased for $\widehat{\mathrm{OR}}_I$. This bias is partly due to the skewed distribution of the estimated odd-ratios, due to the multiplicative structure of the parameter. The error rates were close to the nominal rates in all the cases.

Table 3: Monte Carlo percent RB (in %) and error rates of the Bootstrap variance estimator

	RB	Nominal error rate			
		Lower limit		Upper limit	
		2.5	5.0	2.5	5.0
$\hat{p}_{1\bullet,I}$	-3.9	2.9	5.2	3.4	5.7
$\hat{p}_{\bullet 1,I}$	-5.0	3.4	5.9	3.9	6.4
$\hat{p}_{11,I}$	-3.9	2.5	5.6	3.4	6.1
$\mathrm{OR_I}$	16.2	3.2	5.2	3.3	5.8

7 Concluding Remarks

In this paper, we considered the problem of preserving the relationship between categorical variables when imputation was used to compensate for the missing values. We proposed a simple joint imputation procedure that succeeds in preserving the relationship between two categorical variables, unlike random hot-deck imputation. We also proposed a fully efficient version of the proposed joint imputation procedure. Simulation results showed the good performance of both methods in terms of bias. Also, the balanced joint random hot-deck imputation procedure was

found to be significantly more efficient than the joint random hot-deck imputation procedure.

The properties of the proposed joint imputation procedures were derived under the assumption that the data are MCAR within the imputation classes, which is essentially equivalent to the MAR assumption. An alternative imputation procedure satisfying the MAR assumption consists of using the auxiliary information to obtain a vector of estimated response probabilities. That is, the vector of probabilities $\phi_i \equiv (\phi_{irr}, \phi_{irm}, \phi_{imr}, \phi_{imm})^\top$ may be modeled by using a set of auxiliary variables u_i available for the whole sample. For example, we may assume that this vector of probabilities may be parametrically modeled as

$$\phi_i \;=\; f(u_i;\beta),$$

with $f(\cdot)$ some known function and β a vector of unknown parameters. The estimated probabilities $\hat{\phi}_i = f(u_i;\hat{\beta})$ resulting from replacing β with some estimator $\hat{\beta}$ (for example, obtained through polytomous logistic regression) can then be used to define imputation weights which are used for selecting imputed values to fill-in missing values (see Haziza and Rao, 2006; Chauvet et al., 2016). The theoretical properties of such methods requires further investigations.

An alternative to balanced imputation is fractional imputation, where parameter estimation is performed through the EM algorithm. In the context of fractional imputation, imputed values are assigned a fractional weight; see Kim and Shao (2013, pp. 88) for a discussion of fractional imputation for estimating joint probabilities. An empirical comparison between fractional imputation and the proposed balanced procedure will be presented elsewhere.

In Section 5, we considered the case of negligible sampling fractions and gave an illustration of the use of the Rao, Wu and Yue (1992) bootstrap method for variance estimation. Though numerous bootstrap methods have been proposed in the literature, their drawback is that they are not usually suitable for general sampling designs, in the sense that a particular sampling design usually requires a tailor made resampling scheme. Estimating the variance under the proposed imputation methods with non-negligible sampling fractions and with a general sampling design is a challenging problem. The derivation of linearization variance estimators for the proposed balanced joint random hot-deck imputation procedure is currently under investigation.

Acknowledgments

Part of this work was conducted while Hélène Chaput, Laurianne Salembier and Julie Solard were working at the *Division revenu et patrimoine des ménages* (INSEE). The research of the third author was supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors are thankful to the editors of this special issue, Professor H. Chandra, Professor G. Datta and Professor V.K. Gupta, for inviting us to write a paper in honour of Professor J.N.K. Rao who has made pioneering contributions to the theory and practice of survey sampling.

References

Andridge, R.R. and Little, R.J.A (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, **78**, 40–64.

Antal, E., and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, **106**, 534–543.

Beaumont, J.-F., and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to poisson sampling. *International Statistical Review* **80**, 127–148.

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*, to appear.

Chauvet, G. (2007). Méthodes de Bootstrap en population finie. *Ph.D. dissertation*, University of Rennes 2.

Chauvet, G. (2015). Coupling methods for multistage sampling. *Annals of Statistics*, **43**, 2484–2506.

Chauvet, G., Deville, J.C., and Haziza, D. (2011). On Balanced Random Imputation in Surveys. *Biometrika*, 98, 459–471.

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

Deville, J.C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, **25**, 193–203.

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

Fay, R. E. (1991). A Design-Based Perspective on Missing Data Variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429–440.

Hajek, J. (1971) Comment on An essay on the logical foundations of survey sampling by Basu, D. in Foundations of Statistical Inference (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Haziza, D. and Beaumont, J-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, **75**, 25–43.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of Statistics, Volume 29, Sample Surveys: Theory Methods and Inference*, Editors: C.R. Rao and D. Pfeffermann, 215–246.

Haziza, D., and Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology* **32**, 53–64.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

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

Kim, J.K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Chapman & Hall, CRC. 1

Mashreghi, Z., Léger, C. and Haziza, D. (2014). Bootstrap Methods for Imputed Data from Regression, Ratio and Hot Deck Imputation. *The Canadian Journal of Statistics*, **42**, 162–167.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, 63, 581–590.

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

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**, 254–265.

Shao, J. and Tu, D. (1995). *The jackknife and the bootstrap*. Springer.

Skinner, C. J. and Rao, J. N. K. (2002). Jackknife variance for multivariate statistics under hot deck imputation from common donors. *Journal of Statistical Planning and Inference*, **102**, 149–167.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman and Hall/CRC.

White, I.R., and Daniel, R., and Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics and Data Analysis*, **54**, 2267–2275.

A Proof of (2.6)

From the definition of $\hat{p}_{k\bullet,I}$, we have

$$\begin{aligned} E_I\left(\hat{p}_{k\bullet,I}\right) &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{r\bullet}} w_i 1(x_i=k) \\ &+ N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mr}} w_i\hat{p}^g_{k\bullet,ac} + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mm}} w_i\sum_{l=0}^{L-1}\hat{p}^g_{kl,cc}. \end{aligned}$$

Since $E_q(\hat{p}^g_{k\bullet,ac}) \simeq \hat{p}^g_{k\bullet}$ and $E_q(\sum_{l=0}^{L-1}\hat{p}^g_{kl,cc}) \simeq \hat{p}^g_{k\bullet}$, we obtain

$$\begin{aligned} E_{qI}\left(\hat{p}_{k\bullet,I}\right) &\simeq N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g}w_i(\phi^g_{rr}+\phi^g_{rm})1(x_i=k) \\ &+ N^{-1}\sum_{g=1}^{G}\hat{p}^g_{k\bullet}\sum_{i\in s^g}w_i\phi^g_{mr} + N^{-1}\sum_{g=1}^{G}\hat{p}^g_{k\bullet}\sum_{i\in s^g}w_i\phi^g_{mm} \\ &= N^{-1}\sum_{g=1}^{G}(\phi^g_{rr}+\phi^g_{rm})\sum_{i\in s^g}w_i 1(x_i=k) + N^{-1}\sum_{g=1}^{G}(\phi^g_{mr}+\phi^g_{mm})\sum_{i\in s^g}w_i 1(x_i=k) \\ &= N^{-1}\sum_{i\in s} w_i 1(x_i=k), \end{aligned}$$

so that $B_{qI}\left(\hat{p}_{k\bullet,I}\right) \simeq 0$. The proof for $\hat{p}_{\bullet l,I}$ is similar. We now turn to $\hat{p}_{kl,I}$. From definition, we have

$$\begin{aligned} E_I(\hat{p}_{kl,I}) &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{rr}} w_i 1(x_i=k)1(y_i=l) + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{rm}} w_i 1(x_i=k)\hat{p}^g_{\bullet l,ac} \\ &+ N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mr}} w_i\hat{p}^g_{k\bullet,ac}1(y_i=l) + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mm}} w_i\hat{p}^g_{kl,cc} \\ &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{rr}} w_i 1(x_i=k)1(y_i=l) + N^{-1}\sum_{g=1}^{G}\hat{p}^g_{\bullet l,ac}\sum_{i\in s^g_{rm}} w_i 1(x_i=k) \\ &+ N^{-1}\sum_{g=1}^{G}\hat{p}^g_{k\bullet,ac}\sum_{i\in s^g_{mr}} w_i 1(y_i=l) + N^{-1}\sum_{g=1}^{G}\hat{p}^g_{kl,cc}\sum_{i\in s^g_{mm}} w_i. \end{aligned}$$

Since $E_q(\hat{p}^g_{\bullet l,ac}) \simeq \hat{p}^g_{\bullet l}$, $E_q(\hat{p}^g_{k\bullet,ac}) \simeq \hat{p}^g_{k\bullet}$ and $E_q(\hat{p}^g_{kl,cc}) \simeq \hat{p}^g_{kl}$, we obtain

$$\begin{aligned}
E_{qI}(\hat{p}_{kl,I}) &\simeq N^{-1}\sum_{g=1}^{G}\phi^g_{rr}\sum_{i\in s^g} w_i 1(x_i=k)1(y_i=l) + N^{-1}\sum_{g=1}^{G}\hat{p}^g_{\bullet l}\times\phi^g_{rm}\sum_{i\in s^g} w_i 1(x_i=k)\\
&+ N^{-1}\sum_{g=1}^{G}\hat{p}^g_{k\bullet}\times\phi^g_{mr}\sum_{i\in s^g} w_i 1(y_i=l) + N^{-1}\sum_{g=1}^{G}\hat{p}^g_{kl}\times\phi^g_{mm}\sum_{i\in s^g} w_i\\
&= N^{-1}\sum_{g=1}^{G}(\phi^g_{rr}+\phi^g_{mm})\sum_{i\in s^g} w_i 1(x_i=k)1(y_i=l)\\
&+ N^{-1}\sum_{g=1}^{G}(\hat{N}^g)^{-1}(\phi^g_{rm}+\phi^g_{mr})\left\{\sum_{i\in s^g} w_i 1(x_i=k)\right\}\left\{\sum_{j\in s^g} w_j 1(y_j=l)\right\},
\end{aligned}$$

This leads to

$$\begin{aligned}
E_{qI}(\hat{p}_{kl,I}-\hat{p}_{kl}) &= N^{-1}\sum_{g=1}^{G}(\hat{N}^g)^{-1}(\phi^g_{rm}+\phi^g_{mr})\left\{\sum_{i\in s^g} w_i 1(x_i=k)\right\}\left\{\sum_{j\in s^g} w_j 1(y_j=l)\right\}\\
&- N^{-1}\sum_{g=1}^{G}(\phi^g_{rm}+\phi^g_{mr})\sum_{i\in s^g} w_i 1(x_i=k)1(y_i=l)\\
&= -N^{-1}\sum_{g=1}^{G}(\phi^g_{rm}+\phi^g_{mr})\sum_{i\in s^g} w_i\{1(x_i=k)-\hat{p}^g_{k\bullet}\}\{1(y_i=l)-\hat{p}^g_{\bullet l}\},
\end{aligned}$$

and

$$E_{pqI}(\hat{p}_{kl,I}-\hat{p}_{kl}) \simeq -N^{-1}\sum_{g=1}^{G}(\phi^g_{rm}+\phi^g_{mr})\sum_{i\in U^g}\{1(x_i=k)-p^g_{k\bullet}\}\{1(y_i=l)-p^g_{\bullet l}\},$$

which leads to (2.6).

B Non-response Bias for the Imputed Estimators under the Proposed Procedures

We first consider $\hat{p}_{k\bullet,I}$. From definition, we have

$$\begin{aligned}
E_I\left(\hat{p}_{k\bullet,I}\right) &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{r\bullet}} w_i 1(x_i=k)\\
&+ N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mr}} w_i\sum_{l=0}^{L-1}1(y_i=l)\hat{p}^g_{k|l,cc} + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mm}} w_i\sum_{l=0}^{L-1}\hat{p}^g_{kl,cc}.
\end{aligned} \tag{B.1}$$

Since $E_q(\hat{p}^g_{k|l,cc}) \simeq \frac{\hat{p}^g_{kl}}{\hat{p}^g_{\bullet l}}$ and $E_q(\sum_{l=0}^{L-1} \hat{p}^g_{kl,cc}) \simeq \hat{p}^g_{k\bullet}$, we obtain

$$\begin{aligned} E_{qI}\left(\hat{p}_{k\bullet,I}\right) &\simeq N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g} w_i(\phi^g_{rr}+\phi^g_{rm})1(x_i=k) \\ &+ N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g} w_i\phi^g_{mr}\sum_{l=0}^{L-1}1(y_i=l)\frac{\hat{p}^g_{kl}}{\hat{p}^g_{\bullet l}} + N^{-1}\sum_{g=1}^{G}\hat{p}^g_{k\bullet}\sum_{i\in s^g} w_i\phi^g_{mm} \\ &= N^{-1}\sum_{i\in s} w_i 1(x_i=k) = \hat{p}_{k\bullet}, \end{aligned}$$

so that $B_{pqI}\left(\hat{p}_{k\bullet,I}\right) \simeq 0$. The proof for $\hat{p}_{\bullet l,I}$ is similar. We now turn to $\hat{p}_{kl,I}$. Using similar arguments, we obtain

$$\begin{aligned} E_I(\hat{p}_{kl,I}) &= N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{rr}} w_i 1(x_i=k)1(y_i=l) + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{rm}} w_i 1(x_i=k)\hat{p}^g_{l|k,cc} \\ &+ N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mr}} w_i 1(y_i=l)\hat{p}^g_{k|l,cc} + N^{-1}\sum_{g=1}^{G}\sum_{i\in s^g_{mm}} w_i\hat{p}^g_{kl,cc} \end{aligned} \tag{B.2}$$

and

$$\begin{aligned} E_{qI}(\hat{p}_{kl,I}) &\simeq N^{-1}\sum_{g=1}^{G}\phi^g_{rr}\sum_{i\in s^g} w_i 1(x_i=k)1(y_i=l)) + N^{-1}\sum_{g=1}^{G}\phi^g_{rm}\sum_{i\in s^g} w_i 1(x_i=k)\frac{\hat{p}^g_{kl}}{\hat{p}^g_{k\bullet}} \\ &+ N^{-1}\sum_{g=1}^{G}\phi^g_{mr}\sum_{i\in s^g} w_i 1(y_i=l)\frac{\hat{p}^g_{kl}}{\hat{p}^g_{\bullet l}} + N^{-1}\sum_{g=1}^{G}\phi^g_{mm}\sum_{i\in s^g} w_i\hat{p}^g_{kl} \\ &= N^{-1}\sum_{i\in s} w_i 1(x_i=k)1(y_i=l) = \hat{p}_{kl}, \end{aligned}$$

so that $B_{pqI}\left(\hat{p}_{kl,I}\right) \simeq 0$.

C Extension of the Proposed Imputation Procedures

In this section, we briefly describe the set-up and extension of the imputation procedures to the case of more than two missing items. To avoid intricate notations, we focus on the case of 3 missing items and describe the extension of the joint random hot-deck imputation only. In addition to x and y, let z denote a study variable with Q possible characteristics $z_i = 0, \ldots, Q-1$ for unit i. We want to impute jointly the three variables x, y and z. We assume that the population U is partitioned into G imputation classes $U_1, \ldots, U_G$ and note $s^g_\circ$ the subset of units in $s^g = S \cap U^g$ with pattern $\circ \in \{rrr, mrr, rmr, rrm, mmr, mrm, rmm, mmm\}$, where the first letter in $\circ$ refers to the status of x (respondent or missing), the second to the status of y and the third to the status of z. We assume that the data are MCAR within imputation classes, and we note $P(i \in s^g_\diamond | i \in s) = \phi^g_\diamond$.

The joint random imputation procedure described in Section 3 can be extended by modeling the distribution of each variable conditionally on the non-missing items known for this variable.

For any unit $i \in U^g$; we note

$$\begin{aligned}
\hat{p}^g_{k|lq,cc} &= \frac{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(y_i=l)1(z_i=q)}{\sum_{i\in s^g_{rrr}} w_i 1(y_i=l)1(z_i=q)},\\
\hat{p}^g_{l|kq,cc} &= \frac{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(y_i=l)1(z_i=q)}{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(z_i=q)},\\
\hat{p}^g_{q|kl,cc} &= \frac{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(y_i=l)1(z_i=q)}{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(y_i=l)},
\end{aligned}$$

for the estimated conditional probabilities when two items are available; we note

$$\begin{aligned}
\hat{p}^g_{kl|q,cc} &= \frac{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(y_i=l)1(z_i=q)}{\sum_{i\in s^g_{rrr}} w_i 1(z_i=q)},\\
\hat{p}^g_{kq|l,cc} &= \frac{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(y_i=l)1(z_i=q)}{\sum_{i\in s^g_{rrr}} w_i 1(y_i=l)},\\
\hat{p}^g_{lq|k,cc} &= \frac{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(y_i=l)1(z_i=q)}{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)},
\end{aligned}$$

for the estimated conditional probabilities when one item is available; finally, we note

$$\hat{p}^g_{klq,cc} = \frac{\sum_{i\in s^g_{rrr}} w_i 1(x_i=k)1(y_i=l)1(z_i=q)}{\sum_{i\in s^g_{rrr}} w_i}.$$

The joint random imputation procedure is as follows:

(i) for $i \in s^g_{mrr}$, missing x_i is imputed by $x^*_i = k$ with probability $\hat{p}^g_{k|y_i z_i,cc}$;

(ii) for $i \in s^g_{rmr}$, missing y_i is imputed by $y^*_i = l$ with probability $\hat{p}^g_{l|x_i z_i,cc}$;

(iii) for $i \in s^g_{rrm}$, missing z_i is imputed by $z^*_i = q$ with probability $\hat{p}^g_{q|x_i y_i,cc}$;

(iv) for $i \in s^g_{mmr}$, missing (x_i, y_i) is imputed by $(x^*_i, y^*_i) = (k, l)$ with probability $\hat{p}^g_{kl|z_i,cc}$;

(v) for $i \in s^g_{mrm}$, missing (x_i, z_i) is imputed by $(x^*_i, z^*_i) = (k, q)$ with probability $\hat{p}^g_{kq|y_i,cc}$;

(vi) for $i \in s^g_{rmm}$, missing (y_i, z_i) is imputed by $(y^*_i, z^*_i) = (l, q)$ with probability $\hat{p}^g_{lq|x_i,cc}$;

(vii) for $i \in s^g_{mmm}$, missing (x_i, y_i, z_i) is imputed by $(x^*_i, y^*_i, z^*_i) = (k, l, q)$ with probability $\hat{p}^g_{klq,cc}$.

D Properties of the Balanced Joint Random Hot-deck Imputation Procedure

In this Section, we prove that $\hat{p}_{\diamond,I} = \tilde{p}_{\diamond,I}$ for $\diamond \in \{k'\bullet, \bullet l', k'l'\}$ and any characteristics k' and l'. We first consider $\hat{p}_{k'\bullet,I}$ for $k' = 0, \ldots, K-1$. The case of $\hat{p}_{\bullet l',I}$ for $l' = 0, \ldots, L-1$ may be proved

similarly. Using (B.1), we obtain after some algebra that a sufficient condition for $\hat{p}_{k'\bullet,I} = \tilde{p}_{k'\bullet,I}$ is that for any $g = 1, \ldots, G$:

$$\sum_{i \in s^g_{mr}} w_i 1(x^*_i = k') = \sum_{i \in s^g_{mr}} w_i \hat{p}^g_{k'|y_i,cc}, \tag{D.1}$$

$$\sum_{i \in s^g_{mm}} w_i 1(x^*_i = k') = \sum_{i \in s^g_{mm}} w_i \left(\sum_{l'=0}^{L-1} \hat{p}^g_{k'l',cc} \right). \tag{D.2}$$

In (D.1), the first term may be rewritten as

$$\begin{aligned} \sum_{i \in s^g_{mr}} w_i 1(x^*_i = k') &= \sum_{(i,k) \in s^{g*}_{mr}} w_i 1(k = k') \sum_{l'=0}^{L-1} 1(y_i = l') \\ &= \sum_{(i,k) \in s^{g*}_{mr}} \left(\hat{p}^g_{k|y_i,cc} \right)^{-1} \left\{ \sum_{l'=0}^{L-1} (\mathbf{t}_{ik})_{(k'-1)L+l'} \right\}, \end{aligned}$$

and the second term may be rewritten as

$$\begin{aligned} \sum_{i \in s^g_{mr}} w_i \hat{p}^g_{k'|y_i,cc} &= \sum_{i \in s^g_{mr}} \sum_{k=0}^{K-1} w_i \hat{p}^g_{k|y_i,cc} 1(k = k') \sum_{l'=0}^{L-1} 1(y_i = l') \\ &= \sum_{(i,k) \in U^{g*}_{mr}} \left\{ \sum_{l'=0}^{L-1} (\mathbf{t}_{ik})_{(k'-1)L+l'} \right\} \end{aligned}$$

so that (D.1) follows from (3.1). Similarly, (D.2) follows from (3.3). We now consider $\hat{p}_{k'l',I}$ for $k' = 0, \ldots, K-1$ and $l' = 0, \ldots, L-1$. Using (B.2), we obtain after some algebra that a sufficient condition for $\hat{p}_{k'l',I} = \tilde{p}_{k'l',I}$ is that for any $g = 1, \ldots, G$:

$$\sum_{i \in s^g_{mr}} w_i 1(x^*_i = k') 1(y_i = l') = \sum_{i \in s^g_{mr}} w_i \hat{p}^g_{k'|l',cc} 1(y_i = l'), \tag{D.3}$$

$$\sum_{i \in s^g_{rm}} w_i 1(x_i = k') 1(y^*_i = l') = \sum_{i \in s^g_{rm}} w_i \hat{p}^g_{l'|k',cc} 1(x_i = k'), \tag{D.4}$$

$$\sum_{i \in s^g_{mm}} w_i 1(x^*_i = k') 1(y^*_i = l') = \sum_{i \in s^g_{mm}} w_i \hat{p}^g_{k'l',cc}. \tag{D.5}$$

It is easily proved that (3.1), (3.2) and (3.3) imply (D.3), (D.4) and (D.5), respectively.