Statistica Sinica **34** (2024), 637-655
doi:https://doi.org/10.5705/ss.202021.0331

BALANCED DONOR IMPUTATION HANDLING SWISS CHEESE NONRESPONSE

Esther Eustache*[1], Audrey-Anne Vallée[2] and Yves Tillé[1]

[1] *Université de Neuchâtel and* [2] *Université Laval*

Abstract: The estimator of a parameter of interest can be affected significantly by missing values, which introduce bias and cause additional variability. Swiss cheese nonresponse, also known as nonmonotone nonresponse, is difficult to deal with, because it occurs when each variable of a survey may contain missing values, but without any particular pattern. To reduce the effects of nonresponses, missing values are usually imputed. However, when several variables of a data set need to be imputed, it can be difficult to preserve the distributions of the variables and the relationships between them. In this paper, we propose a new donor imputation method that generalizes the balanced k-nearest neighbor imputation, and is applicable to any configuration of item nonresponses. This new method uses random imputations by donors and is constructed to meet the following requirements. First, all missing values of a unit should be imputed by the same donor. Next, a unit with missing values should be imputed by a neighboring donor. Last, the donors are selected to satisfy some balancing constraints that allows us to decrease the variance of the estimator. The method is divided into two phases. First, we create a stratification by computing a matrix of imputation probabilities using linear programming. Then, we select donors using these imputation probabilities and balanced stratified sampling.

Key words and phrases: Donor imputation, linear programming, nonmonotone nonresponse, random imputation.

1. Introduction

In large-scale surveys, nonresponses are often inevitable. There are two types of nonresponse: unit nonresponse, which occurs when all information is missing for a sampled unit, and item nonresponse, which occurs when some, but not all information is missing for a sampled unit. Missing values can affect the estimators of the parameters of interest significantly by introducing bias and causing additional variability. There are two approaches to reducing such effects: the *imputation model*, in which the missing values are imputed, and the *nonresponse model*, in which the responding units are reweighted to compensate for the nonresponding units. Although we focus on donor imputation methods, we also show (Proposition 2) that the estimators can be presented as a reweighting method.

*Corresponding author.

Nonresponses can be univariate or multivariate. In the first case, nonresponses occurs in only one variable, and we can perform imputation using the other fully observed variables. Although several methods exist for univariate imputation, fractional hot deck imputations (FDHIs) are popular in practice (Kim and Fuller (2004); Fuller and Kim (2005)). Recently, Chen and Haziza (2019) reviewed methods (deterministic and random) for univariate imputation, including multiple and fractional imputations.

In the multivariate case, nonresponses occur in more than one variable of the survey. Here, we need to determine whether nonresponses can appear in all variables, or only in some, and whether or not the nonresponses are monotonic. Monotone nonresponse occurs when the missing values follows a specific pattern in the data set, as in longitudinal studies, where there is attrition.

In the first case, the missing values do not appear in all variables of the data set, and are not monotonic. Several methods have been proposed to deal with this missing pattern (Murray and Reiter (2014); Sang, Kim and Lee (2022)). The most general nonresponse pattern is when nonresponses can occur in all survey variables. Here, the difficulty lies in preserving the distributions of the variables and the relationships between them when replacing the missing values. Hasler, Craiu and Rivest (2018) use grapevine copulas to impute monotone nonresponses, and present an overview of other imputation methods for this pattern.

This work is devoted to methods that can be applied to the most general situation, that is, nonmonotone nonresponse, also known as Swiss cheese nonresponse, which occurs when the survey variables all have missing values, but without a particular pattern. Most existing imputation methods are iterative, because of the presence of nonresponses in all variables. van Buuren (2018) reviewed joint modeling and fully conditional specification (FCS) procedures. An example of these iterative algorithms is a sequence of regression models between the variables developed by Raghunathan et al. (2001). However, Chen (2010) argues that FCS methods may encounter difficulties due to model incompatibilities. Stekhoven and Bühlmann (2011) developed a widely used and efficient iterative imputation method based on random forest models.

Donor imputation methods impute the missing values of a unit using values from other responding units, called donors. The advantage of this method is that the imputed values are plausible, because they are observed for the donor units. Moreover, these methods do not require an iterative system. Yang and Kim (2016) introduced an FHDI for a multivariate nonresponse pattern that is a donor imputation method implemented in the R package FHDI (Im, Cho and Kim (2018)). Judkins (1997) and Andridge and Little (2010) present overviews of donor imputation methods in both univariate and multivariate cases.

Here, we propose a donor imputation method that includes balancing constraints for Swiss cheese nonresponses. This idea of using balancing constraints

for imputing missing values has been considered before. Chauvet, Deville and Haziza (2011) reduced the imputation variance using balanced sampling. Hasler and Tillé (2016) developed a balanced k-nearest neighbor imputation to deal with an univariate nonresponse. This imputation method has the advantage of satisfying balancing equations on the survey variables. Our method extends the balanced k-nearest neighbor imputation to include Swiss cheese nonresponses. This extension is not trivial, because we need to manage missing values for several variables simultaneously and the model cannot be constructed based on completely observed variables.

The proposed imputation method meets three essential requirements. First, in order to preserve the distributions of the variables, it must be a donor imputation method, which allows us to impute continuous and categorical variables using realistic values. Futhermore, all the missing values of a unit should be imputed by the same donor, in order to preserve the relationships between the variables. Second, a unit with missing values must be imputed by a similar donor to ensure consistency between imputed and observed values. Third, we use balancing constraints to reduce the additional variability of the estimated parameters. Note that the proposed method can also be applied to simpler nonresponse patterns, such as monotone or univariate nonresponses.

We present the context and the requirements of the method in Section 2, and the construction of the matrix of imputation probabilities in Section 3. We discuss selecting the donors and the imputation process in Section 4, and the FHDI method in Section 5. In Section 6, we examine several properties of the estimator of the total after imputation using our proposed method. A simulation study using the R package SwissCheese (Eustache, Vallée and Tillé (2021)) is presented in Section 7. Section 8 concludes the paper.

2. Motivations

Consider a finite population U of size N with J variables of interest. A random sample S of size n is selected in U. The first-order inclusion probability of unit i is π_i, the second-order inclusion probability of units i and ℓ is $\pi_{i\ell}$, and $\pi_{ii} = \pi_i$, for any $i, \ell \in U$. The vector of J variables of interest, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ})^\top$, is not necessarily fully observed for all $i \in S$. The vector of response indicators of a unit i is $\mathbf{r}_i = (r_{i1}, \ldots, r_{ij}, \ldots, r_{iJ})^\top$, where r_{ij} is one if the variable j of unit i is observed, and zero otherwise. Consider $S_r \subset S$, the set of $n_r > 0$ units for which the J variables are completely observed. That is, $r_{ij} = 1$, for all $j = 1, \ldots, J$ and any $i \in S_r$. Consider $S_m = S \backslash S_r$, a set of $n_m = (n - n_r)$ units, such that some values, but not all, are missing. Throughout this paper, units in S_r are referred to as respondents, and units in S_m are referred to as nonrespondents. The nonresponse is nonmonotone, and thus has no particular pattern. Figure 1 illustrates a data set with Swiss cheese nonresponses. Note that

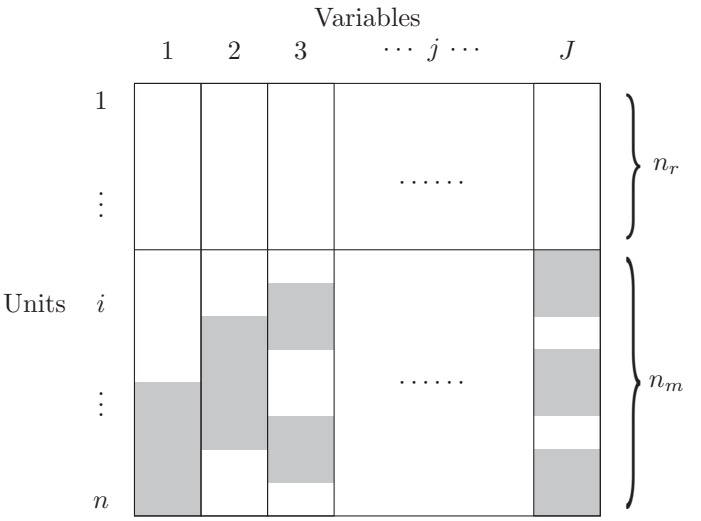

Figure 1. Representation of Swiss cheese nonresponse in a data set of n units and J variables. The first n_r rows correspond to the respondents, and the subsequent n_m rows correspond to nonrespondents. The gray rectangles cover the missing values.

the proposed method can also be applied to simpler configurations, for example, when some variables are not affected by a nonresponse. For example, when a variable j is fully observed, then $r_{ij} = 1$, for all $i \in U$, and the following discussion therefore remains valid.

When no vector $\mathbf{x}_i$ suffers from nonresponse, an unbiased estimator of the population total of the variable j, $T_j = \sum_{i \in U} x_{ij}$, is given by the Horvitz-Thompson estimator

$$\widehat{T}_j^{HT} = \sum_{i \in S} d_i x_{ij},$$

where $d_i = \pi_i^{-1}$ is the sampling weight of unit i. If values are missing in the data set, then they can be imputed, where the imputed value of unit i for a variable j is denoted by x_{ij}^*. Then, T_j is estimated by

$$\widehat{T}_j = \sum_{i \in S} r_{ij} d_i x_{ij} + \sum_{i \in S_m} (1 - r_{ij}) d_i x_{ij}^*.$$

The proposed method ensures coherence and accuracy in the imputed data set, and is based on the following three requirements:

(i) The imputed values should be selected from the values of the n_r fully observed units: a donor imputation method should be used. Furthermore, all missing values of a nonrespondent should be imputed using the same donor.

(ii) The donors should be as close as possible to the nonrespondents, in terms of the distance between survey variables.

(iii) If the observed values of the nonrespondents are imputed, the estimator of the total of each survey variable should remain unchanged.

Requirement (i) ensures that the imputed values are observed, and therefore realistic, for both categorical and continuous variables. Futhermore, a random imputation method tends to preserve the distributions of the variables. To illustrate Requirement (i), consider $J = 3$ variables and a nonrespondent $v \in S_m$, such that $\mathbf{r}_v = (1, 0, 0)^\top$. The missing values of unit v, x_{v2} and x_{v3}, are imputed using observed values selected from the same donor. This means that x_{v2} and x_{v3} are imputed by x_{u2} and x_{u3}, respectively, of a selected donor $u \in S_r$. Requirement (i) aims to preserve the relationships between variables.

Requirement (ii) allows the imputation of a nonrespondent using a similar unit. This ensures coherence between the imputed and the observed values of a nonrespondent. For instance, if we are recording the sex and height of people, the missing height of a man should be imputed using the height of another man. Requirement (ii) also aims to preserve the relationships between variables.

The idea behind Requirement (iii) is that the observed information would remain unchanged if the units with missing values were completely imputed. The estimators based on known values would not be affected. This requirement reduces the variance of the estimators.

To implement a donor imputation method, each fully observed unit receives a probability of donating its values to each nonrespondent. Next, we select one donor per nonrespondent, based on these imputation probabilities. The imputation probabilities satisfying Requirements (i)–(iii) are discussed further in Section 3. The selection of donors is discussed in Section 4.

3. Imputation Probabilities

3.1. Matrix of imputation probabilities

The first step of a donor imputation method is to assign imputation probabilities to the units in the set of respondents. Consider $\boldsymbol{\psi} = (\psi_{uv})$, where $(u, v) \in S_r \times S_m$, the matrix of imputation probabilities. The element ψ_{uv} is the probability that respondent u is the donor selected to impute the missing items of nonrespondent v, with $\psi_{uv} \in [0, 1]$. We need to impose some additional constraints on the imputation probabilities in order to meet Requirements (i)–(iii).

First, only one donor should be randomly selected for a unit $v \in S_m$; see Requirement (i). To this end, if the donors are chosen using balanced sampling, as suggested in Section 4, it is sufficient to ensure that the imputation probabilities

associated with a nonrespondent sum to one; that is

$$\sum_{u \in S_r} \psi_{uv} = 1, \quad v \in S_m. \tag{3.1}$$

Requirement (iii) suggests that if the observed values of any $v \in S_m$ are imputed, the estimator of the total of each variable remains equal to the total of the observed values in S_m. Therefore, the imputation probabilities are chosen so that if the known values of the units in S_m were imputed by the expectation of their imputed values, the estimator of the total would correspond to that based on the observed values. This means that the imputation probabilities must satisfy

$$\sum_{v \in S_m} d_v r_{vj} \sum_{u \in S_r} \psi_{uv} x_{uj} = \sum_{v \in S_m} d_v r_{vj} x_{vj}, \quad j \in \{1, \ldots, J\}, \tag{3.2}$$

see also Figure 2. The right-hand side of Equation (3.2) is the estimated total of the jth variable based on the observed values in S_m; see Figure 2b. The left-hand side of Equation (3.2) is the same estimated total, but calculated using imputed values in S_m. Each observed value x_{vj}, such that $v \in S_m$ and $r_{vj} = 1$, is imputed by

$$x^*_{vj} = \sum_{u \in S_r} \psi_{uv} x_{uj}.$$

The hatched region in Figure 2c represents these values. Then, the total of these imputed values corresponds to the left-hand side of Equation (3.2); see Figure 2c.

Requirement (ii) implies that the donor must be similar to the nonrespondent, where similarity is defined in terms of the distance between units. Let $d(\cdot, \cdot)$ denote a distance function. The closer the distance $d(u, v)$ is to zero, the more similar the units u and v are. After computing the distance between a nonrespondent v and all responding units in S_r, those with the smallest distances to v should have the highest probabilities of being a donor for v.

In other words, for each nonrespondent $v \in S_m$, we want to select the donor $u \in S_r$ that minimizes the product $d(u, v)\psi_{uv}$. For instance, the distance between a respondent u and a nonrespondent v could be the Euclidean distance where the variables with missing values do not contribute to the distance, such that

$$d(u, v) = \left\{ \sum_{j=1}^{J} r_{vj} (x_{uj} - x_{vj})^2 \right\}^{1/2}.$$

The variable must be standardized before calculating the distance because of possible differences in the magnitudes.

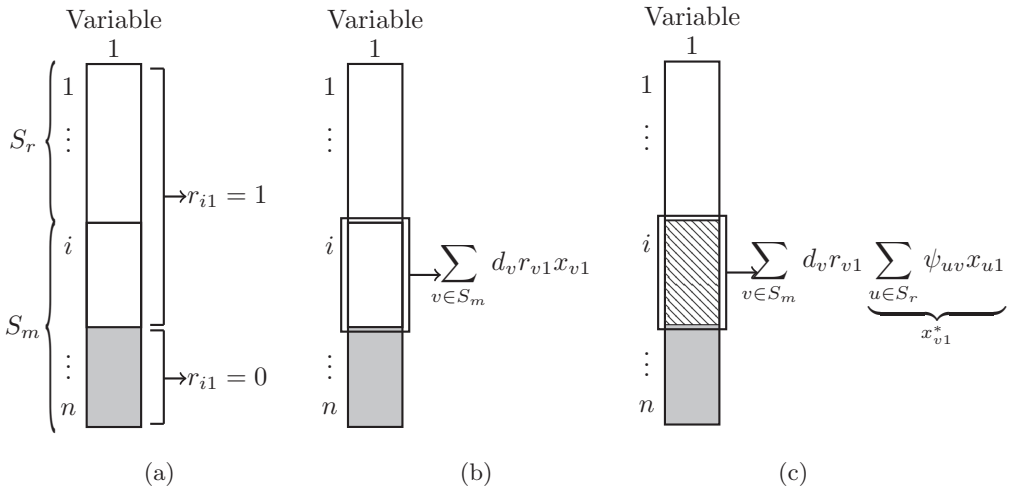

Figure 2. Representation of Swiss cheese nonresponses for the variable $j = 1$. The gray rectangles cover the missing values. Figure 2a represents the variable and the sets S_r and S_m of respondents and nonrespondents, respectively. For unit i with a missing value at variable one, the corresponding response indicator r_{i1} is zero. Figure 2b represents the right-hand side of Equation (3.2) for the variable $j = 1$. In Figure 2c, the observed values in S_m are imputed and represented in the hatched region. The left-hand side of Equation (3.2) is represented, and x^*_{v1} is the imputed value for nonrespondent v.

The matrix $\boldsymbol{\psi}$ satisfying equations (3.1) and (3.2) can be found by solving the linear program

$$\begin{cases} \underset{\psi_{uv} \in [0,1]}{\text{minimize}} \displaystyle\sum_{v \in S_m} \sum_{u \in S_r} d(u,v)\psi_{uv}, & \\ \text{subject to } \displaystyle\sum_{u \in S_r} \psi_{uv} = 1, & v \in S_m, \\ \displaystyle\sum_{v \in S_m} d_v r_{vj} \sum_{u \in S_r} \psi_{uv} x_{uj} = \sum_{v \in S_m} d_v r_{vj} x_{vj}, & j = 1, \ldots, J. \end{cases} \tag{3.3}$$

A solution to (3.3) can almost always be found when the number of respondents n_r is large, because in this case, the constraints are not too restrictive. If the sample size n is small, it is preferable to have at least $n_r/n = 0.5$ to satisfy the balancing constraints and to find similar donors for each nonrespondent.

Consider the bipartite set of S_r and S_m, $U^* = S_r \times S_m$, of size $n_r \cdot n_m$. The calculation of the final imputation probabilities ψ_{uv} can be viewed as a stratification process. A stratum is assigned to each nonrespondent, such that the population U^* is stratified in n_m strata $U^*_v = \{(u,v) | u \in S_r\}$, for $v \in S_m$. Each stratum corresponds to one nonrespondent and contains the set of n_r possible donors for nonrespondent v. Then, a sample of cells must be selected. Each element u, or possible donor, of a stratum U^*_v has a probability ψ_{uv} of being the

selected donor for nonrespondent v. Hence, the inclusion probability of the cell (u, v) is ψ_{uv}, for $(u, v) \in U^*$.

After solving (3.3), in most cases, almost all the probabilities ψ_{uv} obtained are equal to either zero or one. This is equivalent to having a stratum of neighbors consisting of a single respondent. In the next section, we adjust the imputation probability calculation process to enable us to select the minimum number of elements in each stratum.

3.2. The number of neighbors k

After solving (3.3), in most cases, almost all the probabilities ψ_{uv} obtained are equal to either zero or one. However, many researchers encourage considering more than one donor for each non-respondent, for example, as in Jonsson and Wohlin (2004), which adds randomness to the process. This may help to preserve the distribution of the variables and reduce the bias. The constraint that the imputation probabilities need to be smaller than or equal to a quantity k^{-1} can be added to (3.3). Thus, at least k respondents will have a probability greater than zero of being a donor for a nonrespondent v, with $0 < k < n_r$ and $v \in S_m$. The program becomes

$$\begin{cases} \underset{\psi_{uv}\in[0,k^{-1}]}{\text{minimize}} \displaystyle\sum_{v\in S_m}\sum_{u\in S_r} d(u,v)\psi_{uv}, & \\ \text{subject to } \displaystyle\sum_{u\in S_r}\psi_{uv} = 1, & v \in S_m, \\ \displaystyle\sum_{v\in S_m} d_v r_{vj} \sum_{u\in S_r}\psi_{uv}x_{uj} = \sum_{v\in S_m} d_v r_{vj} x_{vj}, & j = 1,\ldots,J. \end{cases} \tag{3.4}$$

The number of neighbors k must be chosen well, because it is used to add randomness to the imputation process. A larger k leads to greater variance due to randomness in the method. We recommend choosing k not greater than five although this depends on the size of the data set and the similarities between the responding units.

4. Imputation

Once we have calculated the matrix of imputation probabilities $\boldsymbol{\psi}$, we can randomly select the donors. Consider $\boldsymbol{\phi} = (\phi_{uv})$, where $(u, v) \in S_r \times S_m$, the imputation matrix. The element ϕ_{uv} is 1 if unit u is selected to donate its values to unit v, and zero otherwise. Only one donor is selected per nonrespondent; thus,

$$\sum_{u\in S_r}\phi_{uv} = 1,$$

for each $v \in S_m$. The matrix $\boldsymbol{\phi}$ must satisfy, at best,

$$\sum_{v\in S_m} d_v r_{vj} \sum_{u\in S_r} \phi_{uv} x_{uj} = \sum_{v\in S_m} d_v r_{vj} \sum_{u\in S_r} \psi_{uv} x_{uj}, \tag{4.1}$$

for each variable $j = 1, \ldots, J$. Therefore, a balanced sampling method is used to select the donors, while satisfying the balancing constraints (4.1). To also ensure that only one donor is selected per nonrespondent, the matrix $\boldsymbol{\phi}$ is generated using stratified balanced sampling (Chauvet (2009); Hasler and Tillé (2014); Jauslin, Eustache and Tillé (2021)).

As explained in Section 3.1, one cell must be selected in each stratum of cells $U_v^* = \{(u, v) \mid u \in S_r\}$. The sample of cells is selected using stratified balanced sampling. Jauslin, Eustache and Tillé (2021) propose a method for selecting a stratified balanced sample when the number of strata is large. If the sum of the inclusion probabilities in each stratum is an integer, the method guarantees the selection of a fixed sample size in each stratum. The size of the sample in a stratum is the sum of the inclusion probabilities of the units in this stratum. The matrix $\boldsymbol{\psi}$ is such that $\sum_{u\in S_r} \psi_{uv} = 1$, for any $v \in S_m$, thus, exactly one cell is selected per stratum, that is, one donor is selected per nonrespondent, and Requirement (i) is exactly satisfied. Moreover, by adding balancing vectors, the method can approximately satisfy (4.1) using the cube method (Deville and Tillé (2004)). The balancing variable of each cell $(u, v) \in S_r \times S_m$ is $d_v r_{vj} \psi_{uv} x_{uj}$. Equation (4.1) might only be approximately satisfied because of the complexity of the balancing problem. Therefore, Requirement (iii) is either exactly or approximately fulfilled. Requirement (ii) is also satisfied, because in the matrix $\boldsymbol{\psi}$, only the closest units of each nonrespondent have non-null imputation probabilities.

The imputation of the data set is based on the matrix $\boldsymbol{\phi}$. The missing value of unit v at variable j, such that $r_{vj} = 0$, is imputed randomly as

$$x_{vj}^* = \sum_{u\in S_r} \phi_{uv} x_{uj}. \tag{4.2}$$

It is also possible to use a deterministic version of the proposed imputation method. The expectation of ϕ_{uv} is used for $(u, v) \in S_r \times S_m$. Then, the missing value x_{vj} is imputed as

$$x_{vj}^* = \sum_{u\in S_r} \psi_{uv} x_{uj}. \tag{4.3}$$

Although this is no longer a donor imputation method, Requirement (iii) is exactly satisfied. In general, the presence of a random component helps to preserve the distribution of the variables, for instance, when estimating a nonlinear estimator as a percentile near or in the tail of the distribution.

5. Comparison with FHDI

The FHDI method is reviewed in Yang and Kim (2016), and is popular in practice. Its steps are similar to those of the proposed imputation method, which is a two-phase stratified sampling. First, a set of imputation cells is formed using all observed values for each variable containing nonresponses. For each cell, the imputation weight, called the fractional weight in the FHDI method, is calculated based on the joint probability of the vector of variables $(\mathbf{x}_1, \ldots, \mathbf{x}_J)$. The calculation of the fractional weights is described in Section 4.1 of Yang and Kim (2016). Second, a hot deck imputation is conducted. Similarly to the proposed imputation method, determining the imputation cells and imputation weights corresponds to stratification, and the hot deck imputation corresponds to stratified sampling. However, although the methods have the same structure, their procedures are different.

FHDI requires discretizing continuous variables to compute the imputation weights. The discretization of each continuous variable is done by dividing its range into a small finite number of segments, as quantiles, for example. This loss of information may become a problem when the number of variables J increases. In addition, the final imputation is a weighted average of the values of the responding units. Thus, the imputed values are not true observed values, but rather a function of several values, and the method is not random. To address this problem, the imputation process described in Section 4 replaces the weights ψ_{uv} with the fractional weights. The FHDI method is considered in the simulation study in Section 7.

6. Properties of the Imputed Estimator of the Total

The proposed imputation method provides a reliable estimation in several different cases. Here, we show that the estimator can be interpreted both as a prediction imputation method and as a reweighting method. Depending on the interpretation, the estimator of the total $\widehat{T}_j$, with the imputed values given in Equation (4.2), can be unbiased, under certain assumptions. In the section, we propose three assumptions that imply unbiasedness. The inference is valid when only one of them is satisfied. Some are on the prediction model, and some are on the weights.

Let $\mathrm{E}_p(.)$, $\mathrm{E}_q(.)$ and $\mathrm{E}_{imp}(.)$ denote the expectation with respect to the sampling design, nonresponse mechanism, and random imputation, respectively. The propositions presented in this section hold only when data are missing at random or completely missing at random, in the sense of Rubin (1976).

Proposition 1. *Consider the notation*

$$\mathbf{x}_i^{(-j)} = (x_{i1}, \ldots, x_{i(j-1)}, x_{i(j+1)}, \ldots, x_{iJ})^\top,$$

for $i \in U$ and $j = 1, \ldots, J$. Suppose further that the context is that of a prediction and assume the following model m:

$$m: \quad x_{ij} = \boldsymbol{\beta}^{(-j)^\top} \mathbf{x}_i^{(-j)} + \varepsilon_i \quad \textit{with} \quad E_m(\varepsilon_i) = 0,$$

where $E_m(.)$ denotes the expectation with respect to the model m. Then, the imputed estimator of the total of the variable j, $\widehat{T}_j$, is unbiased, for $j = 1, \ldots, J$,

$$Bias(\widehat{T}_j) = E_m E_p E_q E_{imp}(\widehat{T}_j - T_j) = 0.$$

The proof is given in the Appendix. Proposition 1 suggests that if a variable $\mathbf{x}_j$ can be explained by a linear combination of the other variables $\mathbf{x}_{g,g\neq j}$, the estimator $\widehat{T}_j$ will be unbiased.

Proposition 2. *The estimator of the total can be viewed as an estimator obtained using a reweighting method, such that*

$$\widehat{T}_j = \sum_{u \in S_r} d_u \left(1 + \pi_u \sum_{v \in S_m} d_v \psi_{uv}\right) x_{uj}.$$

When the weight $\left(1 + \pi_u \sum_{v \in S_m} d_v \psi_{uv}\right)$ is a reasonable approximation of the inverse of the probability of the response, that is when

$$\Pr(u \in S_r | S) \approx \frac{1}{1 + \pi_u \sum_{v \in S_m} d_v \psi_{uv}},$$

then the estimator is approximately unbiased,

$$Bias(\widehat{T}_j) = E_p E_q E_{imp}(\widehat{T}_j - T_j) \approx 0.$$

The proof is given in the Appendix. The estimator of the total can be rewritten as a reweighted estimator, such that

$$\widehat{T}_j = \sum_{u \in S_r} d_u w_u x_{uj}.$$

If the weight w_u is equal to the inverse of the probability of the response to variable j, the estimator is unbiased. In other words, the weight w_u compensates for the nonresponse bias, in the same way that the weight d_i compensates for the sampling bias.

Proposition 3. *The proposed imputation method requires that if $u \in S_r$ is the donor for $v \in S_m$, then $u \in knn(v)$. When*

$$u \in knn(v) \implies (1 - r_{vj})(x_{uj} - x_{vj}) = 0,$$

for all $j = 1, \ldots, J$, then the imputed estimator of the total of the variable j, $\widehat{T}_j$,

is unbiased,

$$Bias(\widehat{T}_j) = E_p E_q E_{imp}(\widehat{T}_j - T_j) = 0.$$

The proof is given in the Appendix. Proposition 3 uses the neighborhood principle. Because each donor is selected in the neighborhood of the recipient, the values of the recipient may be, by definition, close to the values of its donor. The closer the values are, the smaller is the bias of the estimator.

7. Simulation Study

We performed a simulation study to analyze the performance of the proposed imputation methods, using the `R` package `SwissCheese` (Eustache, Vallée and Tillé (2021)). We employ an open-source data set from Johnson (1996) that contains 15 variables of morphological data of $n = 250$ men. Only variables with strong correlations are considered: the waist circumference ($\mathbf{x}_1$), the knee circumference ($\mathbf{x}_2$), the chest circumference ($\mathbf{x}_3$), all three in centimeters, the body density in grams per cubic centimeter ($\mathbf{x}_4$), and the percentage of body fat ($\mathbf{x}_5$).

Swiss cheese nonresponses are generated randomly in the data set $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5)$. Nonresponses are generated for the whole data set, such that no variable is fully observed. For each vector $\mathbf{x}_j$ in which we generated missing values, the nonresponse is non-ignorable, because this is the most difficult type of nonresponse to handle. Define the positive values $g_{ij} = x_{ij} - \min(\mathbf{x}_j)$ and a value α_j such that

$$\sum_{i=1}^{n} \min\left[1, \alpha_j \left(g_{ij} + \frac{\sum_{i=1}^{n} g_{ij}}{n^2}\right)\right] = n_r^{(j)},$$

where $n_r^{(j)}$ is the "expected number of units with a missing value at variable j". The expected value for $n_r^{(j)}$ is 113, which gives a proportion of respondents n_r/n of approximately 45%. The probability p_{ij} that unit i responds to item $j \in \{1, \ldots, 5\}$ is

$$p_{ij} = \min\left[1, \alpha_j \left(g_{ij} + \frac{\sum_{i=1}^{n} g_{ij}}{n^2}\right)\right].$$

Missing values are generated randomly using a uniform variable bounded by zero and one. The response indicator r_{ij} is one if unit i responds to item j, and zero otherwise. When $r_{ij} = 0$, the value x_{ij} is missing.

Eight imputation methods were considered to impute the missing values:

- k-nearest neighbor imputation ($k$nn): a missing value for a nonrespondent is imputed as the mean of this variable for the set of k-nearest neighbors;

- Nearest neighbor (nn): the donor of each nonrespondent is its nearest neighbor;

- FHDI: the imputation method proposed by Yang and Kim (2016) and discussed in Section 5;
- Sequential regression multiple imputation (SReg): an iterative algorithm that imputes variables one by one using a regression model (Raghunathan et al. (2001); van Buuren (2018));
- Balanced nearest neighbors (B-nn): the method proposed in Sections 3 and 4, without constraining a minimum number of neighbors, as in System (3.3), with random imputation as in Equation (4.2);
- Deterministic balanced nearest neighbors (DB-nn): a deterministic version of the B-nn, as in Equation (4.3);
- Balanced k-nearest neighbors (B-$k$nn): the method proposed in Sections 3 and 4, by constraining the minimum number of neighbors to k, as in System (3.4), with random imputation as in Equation (4.2);
- Deterministic balanced k-nearest neighbors (DB-$k$nn): a deterministic version of the B-$k$nn, as in Equation (4.3).

For each method that uses a number of neighbors k (i.e., $k$nn, FHDI, DB-$k$nn, and B-$k$nn), we use $k = 5$. The sequential regression multiple imputation method is a particular case of the fully conditional specification that imputes multivariate missing data on a variable-by-variable basis (van Buuren et al. (2006)). Although the sequential regression multiple imputation is not a donor imputation method, it should work well because of the high correlations between the variables. Based on each imputed data set, we estimate the total of each variable, along with the 50th and the 75th percentiles.

The nonresponse is generated $M_R = 100$ times and, each time, the imputation is repeated $M_I = 100$ times, thus, we create M_R data sets with different nonresponse patterns. For each data set and for each imputation method, we create M_I imputed data sets. Obviously, the M_I imputations for the same nonreponse do not vary for the deterministic methods (i.e., $k$nn, nn, DB-$k$nn, and DB-nn). For each imputation method and parameter, we calculate the Monte Carlo bias of an imputed estimator $\widehat{\theta}$,

$$\text{Bias}\left(\widehat{\theta}\right) = \text{E}_q\text{E}_{imp}\left(\widehat{\theta} - \theta\right) = \frac{1}{M_R M_I}\sum_{r=1}^{M_R}\sum_{i=1}^{M_I}\left(\widehat{\theta}^{r,i} - \theta\right),$$

where $\widehat{\theta}^{r,i}$ is the value of the imputed estimator of the parameter θ in the simulation (r, i), for $r = 1, \ldots, M_R$ and $i = 1, \ldots, M_I$. We also calculate the Monte Carlo mean squared error (MSE) of the imputed estimator,

$$\text{MSE}\left(\widehat{\theta}\right) = \text{E}_q\text{E}_{imp}\left\{\left(\widehat{\theta} - \theta\right)^2\right\} = \frac{1}{M_R M_I}\sum_{r=1}^{M_R}\sum_{i=1}^{M_I}\left(\widehat{\theta}^{r,i} - \theta\right)^2.$$

Table 1. Bias and mean squared errors (MSE) with respect to the imputation and the nonresponse mechanisms, of the estimators of the totals, in the case of $k$nn, nn, FHDI, SReg, DB-nn, B-nn, DB-$k$nn and B-$k$nn imputations. The dataset contains Swiss cheese nonresponse, each variable contains approximately 10% of missing values.

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
True value	*9,083.35*	*9,633.20*	*25,165.50*	*263.96*	*4,757.90*
Bias					
$k$nn	-61.72	-29.88	-106.23	-0.10	-117.36
nn	-59.85	-27.61	-109.99	-0.05	-121.14
FHDI	-10.25	-13.85	-26.41	-0.02	-11.15
SReg	-5.61	-12.39	-16.13	-0.02	-1.59
DB-nn	-8.79	-12.30	-24.25	-0.03	-5.58
B-nn	-8.45	-11.93	-23.36	-0.02	-5.49
DB-$k$nn	-11.62	-12.79	-29.94	-0.02	-6.60
B-$k$nn	-11.33	-12.25	-29.13	-0.01	-6.36
MSE					
$k$nn	4,327.35	1,070.79	14,137.19	0.03	16,193.98
nn	4,712.63	1,122.08	16,226.03	0.04	17,816.19
FHDI	250.57	302.01	1,417.03	0.00	511.15
SReg	111.90	287.71	920.28	0.00	168.77
DB-nn	197.59	259.34	1,675.01	0.01	262.24
B-nn	176.11	242.32	1,556.93	0.00	263.62
DB-$k$nn	219.00	240.36	1,610.32	0.01	198.89
B-$k$nn	247.59	272.59	1,831.69	0.00	370.77

The results for the totals and the percentiles are shown in Tables 1 and 2, respectively, for the eight imputation methods.

For the estimation of the totals, the proposed methods (i.e., B-nn, DB-nn, B-$k$nn, and DB-$k$nn) seem to be equivalent, and outperform the nn and $k$nn imputations in terms of bias and MSE. For the estimation of the percentiles, the proposed methods also outperform than the nn and $k$nn methods. The biases and MSEs of our proposed methods appear to be smaller than those of FHDI for the estimation of the totals. They are similar when estimating a quantile. Globally, the balancing constraints and the donor imputation seem to reduce the bias and MSE of the estimators. The results of our proposed method and those of the SReg imputation are comparable, although the latter is not a donor method. The requirement to use donors is restrictive. Thus, it is promising that our donor methods have almost similar efficiency. Futhermore, the variables are highly correlated, implying that linear regression models are appropriate. SReg is then well suited for the data.

In terms of bias and MSE, the B-nn and DB-nn imputations give similar results, because as expected, almost all the probabilities ψ_{uv} are equal to zero or one, leading to few differences between the two methods. Moreover, they both

Table 2. Bias and mean squared errors (MSE) multiplied by 100 with respect to the imputation and the nonresponse mechanisms, of estimated 50th and 75th percentiles, in the case of $k$nn, nn, FHDI, SReg, DB-nn, B-nn, DB-$k$nn and B-$k$nn imputations. The dataset contains Swiss cheese nonresponse, each variable contains approximately 10% of missing values.

	P50					P75				
	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
True value	*33.28*	*36.92*	*94.25*	*1.04*	*12.42*	*39.04*	*39.87*	*105.30*	*1.07*	*25.20*
Bias										
$k$nn	-15.15	-8.80	-19.11	0.05	-75.75	-43.91	-23.10	-80.67	-0.24	-111.23
nn	-16.77	-7.30	-26.15	0.00	-65.30	-19.25	-10.10	-37.45	-0.04	-55.85
FHDI	0.45	-0.69	-0.78	0.01	-2.34	-6.94	-5.4	-7.36	-0.04	-13.08
SReg	2.37	-2.54	3.14	0.01	-0.98	-4.72	-3.56	3.85	-0.01	-5.17
DB-nn	-0.10	-1.90	-1.55	0.01	-1.20	-1.85	-6.07	-5.88	-0.02	-3.59
B-nn	-0.23	-1.89	-1.62	0.01	-1.34	-1.85	-6.07	-5.97	-0.02	-3.54
DB-$k$nn	1.35	-2.15	1.01	0.01	2.16	-8.46	-7.00	-13.60	0.00	-5.78
B-$k$nn	-0.94	-2.69	-2.84	0.01	-0.34	-3.89	-5.19	-9.92	-0.01	-4.32
MSE										
$k$nn	3.50	1.31	7.90	0.00	67.15	23.07	6.18	76.03	0.00	147.77
nn	3.97	1.12	11.96	0.00	56.67	6.76	1.82	28.27	0.00	42.31
FHDI	0.38	0.41	1.29	0.00	1.89	1.61	0.97	6.12	0.00	5.11
SReg	0.53	0.43	1.37	0.00	0.79	1.04	0.81	5.47	0.00	1.82
DB-nn	0.38	0.46	2.63	0.00	1.21	1.37	0.98	8.32	0.00	2.27
B-nn	0.37	0.48	2.59	0.00	1.23	1.39	0.98	8.43	0.00	2.33
DB-$k$nn	0.36	0.35	1.16	0.00	0.71	1.25	0.88	6.85	0.00	1.57
B-$k$nn	0.53	0.50	2.00	0.00	2.02	1.31	0.91	8.25	0.00	2.62

outperform DB-$k$nn and B-$k$nn. Adding a minimum number $k = 5$ of potential donors to add randomness to the imputation process does not appear to reduce the bias or better preserve the distribution here.

Table 3 shows the bias and MSE of the estimated correlation coefficients between the variables for the B-nn method. The linear relationships between the variables are very well preserved after imputation. We show only the results for the B-nn method, because the other methods yield comparable results.

8. Discussion

In addition to Properties 1–3 on the unbiasedness of the estimated total, the method has two strengths: the possibility of imputing both categorical and continuous variables; and the possibility of forcing the probability ψ_{uv} to be null, if needed, for example, if the survey sampler does not want to allow a respondent u to be the donor of a nonrespondent v.

The variance of estimated parameters is a complex matter when the data sets are imputed, because it needs to consider the variability caused by the

sampling design, nonresponses and the imputation method. Determining an explicit variance estimator requires further investigation, possibly using a pseudo-population bootstrap variance estimator, as described in Chen et al. (2019).

Eustache, Vallée and Tillé (2021) provide a sparse implementation of the methods. The imputation methods can be used in large-scale applications in which both the number of units and the number of variables with missing values are large. With the sparse implementation, the computation of the matrix of imputation probabilities is efficient in terms of computation time.

The choice of the minimum number k of possible donors, as proposed in Section 3.2, depends on the data set. The effect of different values of k on total estimators is left to future research.

Acknowledgments

The authors would like to thank the associate editor and the reviewers for their constructive comments.

Appendix

Proof of Property 1. Consider the column vectors of estimated totals

$$\widehat{\mathbf{T}}_{(-j)} = \left(\widehat{T}_1, \ldots, \widehat{T}_{(j-1)}, \widehat{T}_{(j+1)}, \ldots, \widehat{T}_J\right)^\top$$

and of Horvitz-Thompson estimators

$$\widehat{\mathbf{T}}^{HT}_{(-j)} = \left(\widehat{T}^{HT}_1, \ldots, \widehat{T}^{HT}_{(j-1)}, \widehat{T}^{HT}_{(j+1)}, \ldots, \widehat{T}^{HT}_J\right)^\top,$$

with $\widehat{T}^{HT}_j = \sum_{i\in S} d_i x_{ij}$ the Horvitz-Thompson estimator of the total of variable j. We have that

$$\begin{aligned}
&\mathrm{E}_m\mathrm{E}_{imp}\left(\widehat{T}_j - \widehat{T}^{HT}_j\right)\\
&= \mathrm{E}_m\left(\sum_{i\in S} r_{ij} d_i x_{ij} + \sum_{v\in S_m}(1-r_{vj})d_v \sum_{u\in S_r}\psi_{uv}x_{uj} - \sum_{i\in S} d_i x_{ij}\right)\\
&= \sum_{i\in S} r_{ij} d_i {\boldsymbol{\beta}^{(-j)}}^\top \mathbf{x}_i^{(-j)} + \sum_{v\in S_m}(1-r_{vj})d_v \sum_{u\in S_r}\psi_{uv}{\boldsymbol{\beta}^{(-j)}}^\top \mathbf{x}_u^{(-j)}\\
&\quad - \sum_{i\in S} d_i {\boldsymbol{\beta}^{(-j)}}^\top \mathbf{x}_i^{(-j)}\\
&= {\boldsymbol{\beta}^{(-j)}}^\top\left\{\mathrm{E}_{imp}\left(\widehat{\mathbf{T}}_{(-j)}\right) - \widehat{\mathbf{T}}^{HT}_{(-j)}\right\} = 0.
\end{aligned}$$

The last equality comes from Equation (3.2). Using the requirements that the data are MAR or CMAR, the different expectations can be reversed to obtain

Table 3. Bias and mean squared errors (MSE) with respect to the imputation and the nonresponse mechanisms, of the estimators of the correlation coefficients, in the case of B-nn imputation. The dataset contains Swiss cheese nonresponse, each variable contains approximately 10% of missing values.

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
Bias					
$\mathbf{x}_1$	-	0.0002	-0.0063	-0.0018	-0.0022
$\mathbf{x}_2$		-	0.0001	-0.0058	0.0028
$\mathbf{x}_3$			-	-0.0050	-0.0006
$\mathbf{x}_4$				-	0.0046
$\mathbf{x}_5$					-
MSE					
$\mathbf{x}_1$	-	0.0003	0.0001	0.0001	0.0001
$\mathbf{x}_2$		-	0.0003	0.0006	0.0004
$\mathbf{x}_3$			-	0.0003	0.0001
$\mathbf{x}_4$				-	0.0000
$\mathbf{x}_5$					-

the following development:

$$\begin{aligned}\text{Bias}\left(\widehat{T}_j\right) &= \mathrm{E}_m\mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}(\widehat{T}_j - T_j) = \mathrm{E}_m\mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}(\widehat{T}_j - \widehat{T}_j^{HT} + \widehat{T}_j^{HT} - T_j)\\ &= \mathrm{E}_p\mathrm{E}_q\mathrm{E}_m\mathrm{E}_{imp}(\widehat{T}_j - \widehat{T}_j^{HT}) = 0.\end{aligned}$$

The proof remains the same for each variable $j \in \{1, \ldots, J\}$.

Proof of Property 2.

$$\begin{aligned}\mathrm{E}_{imp}\left(\widehat{T}_j\right) &= \sum_{i\in S} r_{ij}d_i x_{ij} + \sum_{v\in S_m}(1-r_{vj})d_v \sum_{u\in S_r}\psi_{uv}x_{uj}\\ &= \sum_{i\in S_r} d_i x_{ij} + \sum_{\ell\in S_m} r_{\ell j}d_\ell x_{\ell j} + \sum_{v\in S_m}(1-r_{vj})d_v \sum_{u\in S_r}\psi_{uv}x_{uj}\\ &= \sum_{i\in S_r} d_i x_{ij} + \sum_{v\in S_m} d_v \sum_{u\in S_r}\psi_{uv}x_{uj}\\ &= \sum_{i\in S_r} d_i \left\{1 + \pi_i \sum_{v\in S_m} d_v\psi_{iv}\right\} x_{ij}\\ &= \sum_{i\in S_r} d_i w_i x_{ij},\end{aligned}$$

where the third equality comes from Equation (3.2). If w_i^{-1} is approximately equal to the true response probability, we have

$$\text{Bias}(\widehat{T}_j) = \mathrm{E}_p\mathrm{E}_q\mathrm{E}_{imp}\left(\widehat{T}_j - T_j\right) = \mathrm{E}_p\mathrm{E}_q\left(\sum_{i\in S_r} d_i w_i x_{ij} - \sum_{i\in S} x_{ij}\right) \approx 0.$$

Indeed, the quantity

$$\sum_{u\in S_r} \frac{d_u x_{uv}}{\Pr(u \in S_r|S)}$$

is an unbiased estimator of T_j if $\Pr(u \in S_r|S) > 0$, for all $u \in S_r$. If the true response probability is exactly w_i^{-1}, $\widehat{T}_j$ is unbiaised, i.e. $\text{Bias}(\widehat{T}_j) = 0$.

Proof of Property 3.

$$\begin{aligned}
\text{Bias}\left(\widehat{T}_j\right) &= \text{E}_p\text{E}_q\text{E}_{imp}\left(\widehat{T}_j - T_j\right) \\
&= \text{E}_p\text{E}_q\text{E}_{imp}\left(\sum_{i\in S} r_{ij}d_i x_{ij} + \sum_{v\in S_m}(1-r_{vj})d_v \sum_{u\in S_r}\phi_{uv}x_{uj} - \sum_{i\in S} x_{ij}\right) \\
&= \text{E}_p\text{E}_q\text{E}_{imp}\left(\sum_{i\in S} r_{ij}d_i x_{ij} + \sum_{v\in S_m}(1-r_{vj})d_v x_{vj} - \sum_{i\in S} x_{ij}\right) = 0.
\end{aligned}$$

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.

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology* **35**, 115–119.

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

Chen, H. Y. (2010). Compatibility of conditionally specified models. *Statistics & Probability Letters* **80**, 670–677.

Chen, S. and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys: A critical review. *International Statistical Review* **87**, S192–S218.

Chen, S., Haziza, D., Léger, C. and Mashreghi, Z. (2019). Pseudo-population bootstrap methods for imputed survey data. *Biometrika* **106**, 369–384.

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

Eustache, E., Vallée, A.-A. and Tillé, Y. (2021). *The Swisscheese Package*. GitHub Project.

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

Hasler, C., Craiu, R. V. and Rivest, L.-P. (2018). Vine copulas for imputation of monotone non-response. *International Statistical Review* **86**, 488–511.

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

Hasler, C. and Tillé, Y. (2016). Balanced k-nearest neighbor imputation. *Statistics* **50**, 1310–1331.

Im, J., Cho, I. H. and Kim, J. K. (2018). FHDI: An R package for fractional hot deck imputation. *The R Journal* **10**, 140–154.

Jauslin, R., Eustache, E. and Tillé, Y. (2021). Enhanced cube implementation for highly stratified population. *Japanese Journal of Statistics and Data Science* **4**, 783–795.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education* **4**, 1–8. DOI:10.1080/10691898.1996.11910505.

Jonsson, P. and Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using Likert data. In *Proceedings of the 10th International Symposium on Software Metrics*, 108–118. Chicago.

Judkins, D. R. (1997). Imputing for Swiss cheese patterns of missing data. In *Proceedings of Statistics Canada Symposium* **97**, 143–148. Statistics Canada, Ontario.

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

Murray, J. and Reiter, J. P. (2014). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association* **111**, 1466–1479.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85–95.

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

Sang, H., Kim, J. K. and Lee, D. (2022). Semiparametric fractional imputation using Gaussian mixture models for handling multivariate missing data. *Journal of the American Statistical Association* **117**, 654–663.

Stekhoven, D. J. and Bühlmann, P. (2011). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118.

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Boca Raton.

van Buuren, S., Brand, J., Groothuis-Oudshoorn, C. G. and Rubin, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76**, 1049–1064.

Yang, S. and Kim, J. K. (2016). Fractional imputation in survey sampling: A comparative review. *Statistical Science* **31**, 415–432.

Esther Eustache

Institut de Statistique, Université de Neuchâtel, 2000 Neuchâtel, Switzerland.

E-mail: esther.eustache@unine.ch

Audrey-Anne Vallée

Département de mathématiques et de statistique, Université Laval, G1V 0A6 Québec, Canada.

E-mail: Audrey-Anne.Vallee@mat.ulaval.ca

Yves Tillé

Institut de Statistique, Université de Neuchâtel, 2000 Neuchâtel, Switzerland.

E-mail: yves.tille@unine.ch

(Received July 2021; accepted July 2022)