Statistical Methods & Applications (2020) 29:25–48
https://doi.org/10.1007/s10260-019-00458-w

ORIGINAL PAPER

Nonparametric imputation method for nonresponse in surveys

Caren Hasler[1,2] · **Radu V. Craiu**[3]

Accepted: 23 March 2019 / Published online: 4 April 2019

Abstract
Many imputation methods are based on a statistical model that assumes the variable of interest is a noisy observation of a function of the auxiliary variables or covariates. Misspecification of this function may lead to severe errors in estimation and to misleading conclusions. Imputation techniques can therefore benefit from flexible formulations that can capture a wide range of patterns. We consider the use of smoothing splines within an additive model framework to estimate the functional dependence between the variable of interest and the auxiliary variables. The estimator obtained allows us to build an imputation model in the case of multiple auxiliary variables. The performance of our method is assessed via numerical experiments involving simulated and real data.

Keywords Additive models · Data imputation · Sample survey · Smoothing spline

1 Introduction

Nonresponse in surveys is a commonly encountered problem that, when ignored, can affect the performance of the statistical estimators for the quantities of interest. Two general adjustment techniques that have been developed to alleviate the effects of nonresponse are *reweighting* and *imputation*. Reweighting procedures consist of increasing the initial weights of respondents in order to compensate for nonrespondents

Caren Hasler's address when the research was conducted is "Institute of Statistics, University of Neuchâtel, Av. de Bellevaux 51, 2000 Neuchâtel, Switzerland".

✉ Caren Hasler
caren.hasler@utoronto.ca

1 Institute of Statistics, University of Neuchâtel, Av. de Bellevaux 51, 2000 Neuchâtel, Switzerland

2 Present Address: Department of Computer and Mathematical Sciences, University of Toronto Scarborough, 1265 Military Trail, Toronto, ON M1C 1A4, Canada

3 Department of Statistical Sciences, University of Toronto, 100 St. Georges Street, Toronto, ON M5S 3G3, Canada

and are commonly used to treat unit nonresponse, meaning that all the quantities of interest are missing. Imputation procedures consist of filling in the missing values in the data with *imputed values* and are commonly used to treat item nonresponse, i.e. situations in which only some of the quantities of interest are missing. When dealing with nonresponse, both reweighting and imputation may rely on a statistical model. Imputation for the variable of interest can be more efficient if it is based on information contained in a number of auxiliary variables, specifically, through a model that estimates a functional link between the latter and the variable of interest. However, the validity of the model will have a direct effect on the accuracy of the estimated quantities. It is therefore crucial to be able to build flexible models that can capture a large spectrum of patterns and make only weak assumptions about the true underlying mechanism generating the data. Given these constraints, it is not surprising that nonparametric models have been used to handle nonresponse in surveys.

Giommi (1987) focused on unit nonresponse and proposed two nonparametric reweighting procedures based on kernel density estimators to estimate response probabilities. Later, Niyonsenga (1994, 1997) used the nonparametric estimation of Giommi (1987) to handle nonresponse when unit nonresponse and item nonresponse occur together. Finally, Da Silva and Opsomer (2006, 2009) applied, respectively, kernel regression and local polynomial regression to estimate the response probabilities and derived asymptotic properties of the propensity score adjusted estimator for these approaches. These techniques are suitable when the number of auxiliary variables is relatively low. We refer to Ning and Cheng (2012) for a comparison study of nonparametric imputation methods and to Haziza (2009) for a review of imputation and inference with missing data.

We propose here an imputation method for item nonresponse in surveys when the variable of interest is a noisy observation of a function of many auxiliary variables. We consider smoothing spline models within an additive regression framework which allows us to handle a large number of auxiliary variables. This improvement significantly expands the range of nonparametric methods for handling nonresponse. Moreover, the model considered is adaptable to a wide variety of functional patterns thus providing protection against model misspecification. Results of a simulation study confirm the performance of our method and highlight its capacity to adapt to many different situations.

The paper is organized as follows: Sect. 2 establishes the framework and introduces notation; Sect. 3 provides a motivation for the new imputation method; two nonparametric tools used in the new imputation method are reviewed in Sect. 4; Sect. 5 presents the new method as well as bootstrap procedures to estimate the variance of the total. The performance of the new method is compared to that of other imputation methods through a simulation study presented in Sect. 6. We close with concluding remarks and a discussion of future work.

2 Framework

Consider a finite population $U = \{1, 2, \ldots, N\}$ of possibly unknown size N. Suppose that the parameter of interest is the population total

$$Y = \sum_{i \in U} y_i,$$

for some unknown variable of interest y. A sample S of size n is selected from U according to a probabilistic sampling design $p(\cdot)$ with the aim of observing y_i for $i \in S$. We suppose non-informative sampling. We refer the reader to Qin et al. (2002) and Berg et al. (2016) for more details about informative sampling and imputation. Consider

$$\pi_i = \Pr(i \in S) = \sum_{s \subset U; s \ni i} p(s),$$

the first-order inclusion probability of unit i and suppose that $\pi_i > 0$ for all $i \in U$. Let $d_i = 1/\pi_i$ represent the design weight of unit $i \in U$. In this paper we consider two widely used sampling designs, simple random sampling without replacement (SRSWOR) and stratified sampling . While the proposed imputation method can be applied under a general sampling design, the variance estimation requires customization. For the two sampling designs considered here we detail procedures for variance estimation. Under SRSWOR, each sample of (fixed) size n has the same probability of being selected and $\pi_i = n/N$ for all $i \in U$. Under stratified sampling, the population U is partitioned into H strata $U_1, \ldots, U_H$ of respective sizes $N_1, \ldots, N_H$ and SRSWOR is applied independently in each stratum h. A sample S_h of size n_h is hence selected in each stratum U_h, $h = 1, \ldots, H$ and $\pi_i = n_h/N_h$ for all $i \in U_h$.

We assume item nonresponse in which only one variable of interest is either observed or missing. Specifically, unit i in the given sample S is classified as either respondent or nonrespondent, depending on whether y_i is observed or missing. Consider the response indicator vector $(r_i | i \in S)^\top$ where r_i takes value 1 if y_i is observed and 0 if it is missing. This results in the set of respondents $S_r = \{i \in S | r_i = 1\}$ and in the set of nonrespondents $S_m = \{i \in S | r_i = 0\}$. We assume that the missing data are missing at random (see Rubin 1976, for a detailed definition) and that the units' responses are independent of one another.

Under complete response, the Horvitz–Thompson estimator

$$\widehat{Y} = \sum_{i \in S} \frac{1}{\pi_i} y_i, \tag{1}$$

is design-unbiased for Y, i.e. $\mathrm{E}_p(\widehat{Y}) = Y$. In the case of a survey with nonresponse, however, the estimator (1) cannot be computed since some of the y_i's, $i \in S$ are missing. One remedy is to impute each missing value y_i, $i \in S_m$ with an imputed value y_i^*. The population total Y can then be estimated through the *imputed estimator*

$$\widehat{Y}_I = \sum_{i \in S} \frac{1}{\pi_i} \left[y_i r_i + y_i^* (1 - r_i) \right] = \sum_{i \in S_r} \frac{1}{\pi_i} y_i + \sum_{i \in S_m} \frac{1}{\pi_i} y_i^*. \tag{2}$$

Design weights can optionally be taken into account when constructing the imputed values, the resulting method being referred to as *survey weighted imputation*. We refer

the reader to Andridge and Little (2010) and Haziza and Rao (2005) for other examples of survey weighted imputation methods.

We assume an additive imputation model in order to predict the missing values and adopt the imputation model approach (Haziza 2009). With this approach, inference is made with respect to the imputation model, the sampling design, and the nonresponse mechanism. For deterministic imputation, the imputed estimator in (2) is asymptotically unbiased (i.e. the bias of the imputed estimator is negligible as compared to the value of this estimator when the sample size increases to infinity) if the imputation model is correctly specified, the sampling is non-informative, and the data are missing at random (MAR).

3 Motivation

We consider a variable of interest, y, that is measured along with a q-dimensional vector of auxiliary variables, $\mathbf{x} = \left(x_1, x_2, \ldots, x_q\right)^\top$. We assume that $\mathbf{x}_i$ is known for all $i \in S$. We note that auxiliary information can be used at different stages of the survey, namely in establishing the sampling design, for estimation, and handling of nonresponse. Reliable auxiliary information can explain the variation in the variable of interest and/or in the response probabilities and helps reduce error due to sampling and nonresponse.

More importantly for the purpose of this study, in situations in which the variable of interest is not recorded for some sampled units, one may rely on the auxiliary variables to impute the missing values if there is a way to connect these variables via an *imputation model* (Särndal 1992). For instance, consider a general model of the type

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{iq}) + \varepsilon_i, \tag{3}$$

where f is a function from $\mathbb{R}^q$ to $\mathbb{R}$, and ε_i are zero-mean independent errors with variance σ^2. A deterministic imputation method estimates first the function f based on those individuals/items $i \in S_r$ for which $(y_i, \mathbf{x}_i) = (y_i, x_{i1}, \ldots, x_{iq})$ are fully observed, and then imputes values for $i \in S_m$ using the estimated function and the observed $\mathbf{x}_i$. The challenging issue of estimating f arises naturally because the choice of the imputation model crucially impacts the accuracy of the imputed values. A misspecified model may result in highly biased estimates for the parameters of interest.

Without prior knowledge on the form of f in (3), it is natural to use a nonparametric regression model since the resulting estimate $\hat{f}$ is known to adapt to the shape of f based on the information provided by the data. When handling survey data, however, several auxiliary variables are often available and one needs to include most of them in the model. Unfortunately, a few nonparametric smoothers such as kernel-based ones tend to break down in high dimension, unless the sample size is very large. This phenomenon is known as the *curse of dimensionality* (Bellman 1961; Stones 1985) and can be alleviated if an additive model (AM, Hastie and Tibshirani 1986) is used. Such a model is additive in the predictor variables and takes the form

$$y_i = a_0 + \sum_{j=1}^{q} a_j(x_{ij}) + \varepsilon_i, \tag{4}$$

where $(y_i, \mathbf{x}_i) = (y_i, x_{i1}, \ldots, x_{iq})$, $i = 1, \ldots, N$, are observations, a_0 is a constant, a_j, $j = 1, \ldots, q$, are univariate smooth functions, and ε_i are zero-mean independent errors with common variance σ^2. The functions a_j, $j = 1, \ldots, q$, are each individually estimated by univariate smoothers so the curse of dimensionality is avoided because the original problem of nonparametric estimation in $\mathbb{R}^q$ has been replaced by q estimation problems in $\mathbb{R}$. Without loss of generality, henceforth we suppose that the $\mathbf{x}_i$, $i = 1, \ldots, N$, lie in the interval $[0, 1]^q$.

We propose an imputation method for nonresponse in surveys based on AM. The new method is based on imputation model (4). The nonparametric tools used to estimate the regression function are presented in Sect. 4 and the new method is presented in Sect. 5.

4 Nonparametric tools

This section introduces two nonparametric tools used in the new imputation method, smoothing spline regression and additive models. The main idea of smoothing spline regression is to fit a data set with a curve that maximizes a measure of goodness-of-fit while achieving a fixed degree of smoothness. There is an extensive literature devoted to spline regression and we refer the reader to Green and Silverman (1994), Eubank (1999) and Wang (2011). Smoothing spline regression (SSR) assumes model (4) with a unique predictor variable, that is

$$y_i = a(x_i) + \varepsilon_i, \quad 1 \leq i \leq N$$

where ε_i are zero-mean independent errors with common variance σ^2, and a is a smooth function in the sense that $a \in W_2^m[0, 1]$ where $W_2^m[0, 1]$ is the Sobolev space

$$W_2^m[0, 1] = \left\{ g : g, g', \ldots, g^{(m-1)} \text{ are absolutely continuous, } \int_0^1 g^{(m)}(t)^2 dt < +\infty \right\}.$$

We consider a K-dimensional function basis $\{b_k : k \in 1, \ldots K\}$ for the space $W_2^m[0, 1]$. Hence, any function $f \in W_2^m[0, 1]$ can be expressed as

$$f(x) = \sum_{k=1}^{K} \beta_k b_k(x),$$

for some vector of parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)^\top$. Given a function a, SSR aims at finding its best approximation, $\widehat{a} \in W_2^m[0, 1]$ while simultaneously controlling its degree of smoothness. The resulting *smoothing spline estimator* $\widehat{a}$ of a is the minimizer of the following penalized least square (PLS) criterion

$$\widehat{a} = \arg \min_{g \in W_2^m[0,1]} \frac{1}{N} \sum_{i=1}^{N} (y_i - g(x_i))^2 + \lambda \int_0^1 g^{(m)}(t)^2 dt. \tag{5}$$

We consider $m = 2$ in what follows and obtain the following smoothing spline estimator

$$\widehat{a}(x) = \sum_{k=1}^{K} \widehat{\beta}_k b_k(x),$$

where

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{N}^\top \mathbf{N} + N\lambda \boldsymbol{\Omega})^{-1} \mathbf{N}^\top \mathbf{y}, \\ N(i, j) &= b_j(x_i), \\ \Omega(i, j) &= \int_0^1 b_i''(t) b_j''(t) dt. \end{aligned}$$

One may use different basis, each of which yielding a different smoothing spline estimator. In what follows, we will consider the thin plate spline basis (see Wood 2003). The parameter λ is the *smoothing parameter* and its size determines the balance between goodness-of-fit, as measured by the mean squared residual, and smoothness, as measured by the squared L^2 norm of the mth order derivative. The reader may refer to Lee (2003) for a discussion on smoothing parameter selection.

With survey data, it is often desirable to consider design weights when estimating parameters of interest. Indeed, a design weight $d_i = 1/\pi_i$ can be interpreted as the number of population units represented by the ith sampled unit. Hence, when units are selected with unequal inclusion probabilities it might be unreasonable to assume that each sampled unit has the same influence on the parameters of interest. A weighted version of the smoothing spline estimator was proposed by Zhang et al. (2013) who suggested adding design weights in the general PLS criterion in Eq. (5). Hence, they consider the smoothing spline estimator $\widehat{a}_W$ adapted for survey data which is the minimizer of

$$\widehat{a}_W = \arg \min_{g \in W_2^m[0,1]} \frac{1}{\widehat{N}} \sum_{i \in S} d_i \, (y_i - g(x_i))^2 + \lambda \int_0^1 g^{(m)}(t)^2 dt, \tag{6}$$

where $\widehat{N} = \sum_{i \in S} d_i$ is the estimated population size. Note that Zhang et al. (2013) restrict themselves to the case $m = 2$.

A flexible way to combine the contributions of each auxiliary variable to the variable of interest is provided by the additive model paradigm. A class of generalized additive models was proposed by Hastie and Tibshirani (1986) and was discussed in depth in Hastie and Tibshirani (1990). We focus here on the additive regression model (AM) in (4). SSR is used to estimate each function a_j, $j = 1, \ldots, q$. A backfitting algorithm (Hastie and Tibshirani 1986) or a direct fitting approach (Wood 2008) can be considered. The general steps of the backfitting algorithm are:

(1) set initial values for $a_0, a_1, \ldots, a_q$;
(2) keep $a_0, a_2, \ldots, a_q$ constants and estimate a_1 via SSR with model $y_i - a_0 - \sum_{j=2}^{q} a_j(x_{ij}) = a_1(x_{i1}) + \varepsilon_i$;
(3) repeat (2) for $a_2, a_3, \ldots, a_q$;
(4) repeat (2) and (3) until a convergence criterion is met.

Note that it may be required to center the estimated functions, see Hastie and Tibshirani (1986) for more details. The direct fitting approach solves the penalized likelihood maximization problem using penalized iteratively reweighted least squares and is implemented in the `R` package `mgcv` (Wood 2014). When appropriate, an additive model allows us to handle multiple predictor variables in a reasonable computation time and avoids the curse of dimensionality problem as it breaks a high-dimensional nonparametric estimation problem into a number of one-dimensional ones.

5 The method

In this section, we propose a nonparametric model-based imputation method for nonresponse in surveys and discuss bootstrap procedures to estimate the resulting variance of the total estimator for the population U.

5.1 Estimation and imputation

We consider the additive imputation model based on (4). Smoothing spline estimates $\widehat{a}_j$, $j = 1, \ldots, q$, of functions a_j, $j = 1, \ldots, q$, and an estimate $\widehat{a}_0$ of a_0 are obtained using the complete data $(y_i, \mathbf{x}_i)$, $i \in S_r$. Two different smoothing splines estimators can be obtained based on expression (5) (unweighted imputation) or expression (6) (survey weighted imputation), respectively. Finally, missing values y_i, $i \in S_m$, are imputed with predictions based on the imputation model as follows

$$y_i^* = \widehat{a}_0 + \sum_{j=1}^{q} \widehat{a}_j(x_{ij}). \tag{7}$$

The proposed imputation procedure is deterministic. Random imputation methods are preferred when the goal is to estimate population quantiles because they tend to preserve the distribution of the variables being imputed. A simple extension of the imputation scheme in (7) to obtain random imputation consists of adding a random residual to the imputed values. That is, we set

$$y_i^* = \widehat{a}_0 + \sum_{j=1}^{q} \widehat{a}_j(x_{ij}) + e_i^*,$$

where e_i^* is selected at random from the respondent residuals

$$\left\{ y_i - \widehat{a}_0 - \sum_{j=1}^{q} \widehat{a}_j(x_{ij}) : \ i \in S_r \right\},$$

or from the empirical distribution function of the respondent residuals. See Chauvet et al. (2011) on the selection of the random residuals e_i^*.

5.2 Variance estimation for the imputed total

A valid method for estimating the variance of the estimator of the population total must account for the extra variability due to imputing the missing values. In turn, this variability is due to the variance of predicted values y_i^* produced via the additive model. Since an analytical expression for the asymptotic error of AM predictive value is not available, we pursue a bootstrap-based approach. Bootstrap procedures to estimate the variance of parameters of interest are available for different imputation methods and sampling designs. In this Section, we follow Shao and Sitter (1996) to devise bootstrap procedures to estimate the variance of the total under AM imputation for simple random sampling without replacement (SRSWOR) and stratified sampling . The bootstrap variance estimate proposed in Shao and Sitter (1996) is consistent when the sampling fraction is negligible. We refer to Mashreghi et al. (2014) for more details and for bootstrap methods that are valid under three common imputation methods when the sampling fraction is large.

We follow Shao and Sitter (1996) and apply the following methods to estimate the variance of the total under AM imputation: the mirror-match bootstrap (MMB) proposed by Sitter (1992b) and the without-replacement bootstrap (BWO) proposed by Gross (1980). Procedure 1 presents the MMB variance estimation in the case of stratified sampling. For simple random sampling, the same procedure is applied with a single stratum. In step 1, the user chooses a sample size n'_h and selects a simple random sampling of size n'_h from S_h. In step 2, the sample selection procedure of step 1 is repeated k_h times independently. In most cases, k_h is non-integer and a randomization must be applied (see Sitter 1992b, Section 1.6). Note that the randomization may impair the performance of the variance estimator, as detailed in the "Appendix". In step 3, steps 1 and 2 are repeated independently for each stratum h. A bootstrap sample $\mathbf{S}^*$ is obtained. Because the bootstrap sample consists of sampled units, it is very likely to contain units with missing y_i and units with observed y_i. Hence, in step 4, AM imputation is applied to the bootstrap sample $\mathbf{S}^*$ and the bootstrap analog $\widehat{Y}_I^{(b)}$ of the imputed total estimator $\widehat{Y}_I$ is obtained. Depending on the choice of n'_h and on whether randomization is applied to round k_h, the bootstrap procedure might mimic a stratified sampling in a population whose size differs from N. Fraction N/N^* appears in the computation of the bootstrap analog of the imputed total estimator $\widehat{Y}_I^{(b)}$ to take this into account. Steps 1 to 4 are repeated to obtain B analogs of the imputed total estimator. In step 6, the bootstrap variance of the imputed total is obtained using the standard bootstrap formulae. The computational time involved in the bootstrap evaluation of variance can be shortened if multiple processors are available. The embarrassing parallel structure of the procedure implies that the sample-specific calculation can be

performed on a separate processor and the merging of simulated values is needed only in Step 6.

Procedure 1 Variance of the imputed total estimator using MMB.

Step 1: Choose $1 \leq n'_h < n_h$ and select a SRSWOR of size n'_h without replacement from S_h.

Step 2: Repeat step 1 $k_h = n_h(1 - f_h^*)/(n'_h(1 - f_h))$ times independently to obtain a sample $S_h^* = \{hi : i = 1, \ldots, n_h^*\}$ of size $n_h^* = n'_h k_h$, where $f_h = n_h/N_h$ and $f_h^* = n'_h/n_h$.
If k_h is not integer, apply a randomization (see Sitter 1992b, Section 1.6).

Step 3: Repeat steps 1 and 2 independently for each stratum h to obtain a bootstrap sample $\mathbf{S}^* = \{S_1^*, \ldots, S_H^*\} = \{hi : h = 1, \ldots, H; i = 1, \ldots, n_h^*\}$ of size $n^* = \sum_{h=1}^H n_h^*$.

Step 4: Apply AM imputation to impute the bootstrap sample $\mathbf{S}^*$ and obtain the bootstrap analog of the imputed total estimator $\widehat{Y}_I$ by

$$\widehat{Y}_I^{(b)} = \frac{N}{N^*} \sum_{h=1}^{H} \sum_{hi \in S_h^*} \frac{\widetilde{y}_{hi}^{(*)}}{f_h^*} = \frac{N}{N^*} \sum_{h=1}^{H} \frac{n_h}{n'_h} \sum_{hi \in S_h^*} \widetilde{y}_{hi}^{(*)},$$

where $N^* = \sum_{h=1}^H n_h k_h$ and $\widetilde{y}_{hi}^{(*)}$ is the value of the variable of interest of unit hi if this one is observed and the imputed value otherwise.

Step 5: Repeat steps 1 to 4 a large number of times B to obtain $\widehat{Y}_I^{(1)}, \ldots, \widehat{Y}_I^{(B)}$ where $\widehat{Y}_I^{(b)}$ is the analog of $\widehat{Y}_I$ for the b-th bootstrap sample.

Step 6: Obtain the bootstrap variance of $\widehat{Y}_I$ by

$$\mathrm{V}_{boot}(\widehat{Y}_I) = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{Y}_I^{(b)} - \widehat{Y}_I^{(.)}\right)^2,$$

where $\widehat{Y}_I^{(.)}$ is the mean bootstrap analog of $\widehat{Y}_I$

$$\widehat{Y}_I^{(.)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{Y}_I^{(b)}.$$

Procedure 2 presents the variance estimation based on the BWO of Gross (1980) in the case of SRSWOR. Given a sample of size n from a population of size N, we set $k = N/n$ and assume k is an integer. The case of a non-integer $k = N/n$ is discussed in the next paragraph. In step 1 we construct a pseudopopulation of size N by replicating the sample k times. In step 2, a simple random sample of size n is selected from the pseudopopulation. Because the pseudopopulation consists of sampled units, the bootstrap sample is very likely to contain both units with missing y_i and units with observed y_i. In step 3, AM imputation is applied to the bootstrap sample. Steps 2 and 3 are repeated to obtain B analogs of the imputed total estimator. In step 5, the bootstrap variance of the imputed total is obtained using the standard bootstrap formulae.

Procedure 2 can be applied only in the case of simple random sampling and when $k = N/n$ is an integer. We apply the extension of BWO of Sitter (1992a) to stratified sampling and obtain Procedure 3 to estimate the variance under stratified sampling or under simple random sampling when k is a non-integer. For simple random sampling,

Procedure 2 Variance of the imputed total estimator under SRSWOR using BWO when $k = N/n$ is an integer.

Step 1: Consider $k = N/n$. Construct a pseudopopulation of size N by replicating the sample k times.
Step 2: Draw a SRSWOR of size n from the pseudopopulation of step 1. Denote $\mathbf{S}'$ the selected bootstrap sample.
Step 3: Apply AM imputation to impute the sample selected in step 2 and obtain the bootstrap analog of the imputed total estimator $\widehat{Y}_I$ by

$$\widehat{Y}_I^{(b)} = \frac{N}{n} \sum_{i \in S'} \widetilde{y}_i^{(*)},$$

where $\widetilde{y}_i^{(*)}$ is the value of the variable of interest of unit i if this one is observed and the imputed value otherwise.
Step 4: Repeat steps 2 and 3 a large number of times B to obtain $\widehat{Y}_I^{(1)}, \ldots, \widehat{Y}_I^{(B)}$ where $\widehat{Y}_I^{(b)}$ is the analog of $\widehat{Y}_I$ for the b-th bootstrap sample.
Step 5: Obtain the bootstrap variance of $\widehat{Y}_I$ by

$$\mathrm{V}_{boot}(\widehat{Y}_I) = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{Y}_I^{(b)} - \widehat{Y}_I^{(.)} \right)^2,$$

where $\widehat{Y}_I^{(.)}$ is the mean bootstrap analog of $\widehat{Y}_I$

$$\widehat{Y}_I^{(.)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{Y}_I^{(b)}.$$

a single stratum is considered. In step 1, the user computes a sample size n'_h and a replication factor k_h for stratum h. In most cases, at least one of k_h and n'_h is non-integer and a randomization must be applied (see Sitter 1992a, Section 3.1). Note that the randomization may impair the performance of the variance estimator, as detailed in the "Appendix". In step 2, the units in stratum h are replicated k_h times to obtain the h-th stratum of the pseudopopulation. In step 3, a SRSWOR of size n'_h is selected without replacement from the h-th stratum of the pseudopopulation created in step 2. Steps 1 to 3 are repeated independently for each stratum h and a bootstrap sample $\mathbf{S}^*$ is obtained in step 4. Because the bootstrap sample consists of sampled units, it is very likely to contain units with missing y_i and units with observed y_i. Hence, in step 5, AM imputation is applied to the bootstrap sample $\mathbf{S}^*$ and the bootstrap analog $\widehat{Y}_I^{(b)}$ of the imputed total estimator $\widehat{Y}_I$ is obtained. Depending on the choice of k_h and n'_h and on whether randomization is applied to round these quantities, the bootstrap procedure might mimic a stratified sampling in a population whose size differs from N. The fraction N/N^* appears in the computation of the bootstrap analog of the imputed total estimator $\widehat{Y}_I^{(b)}$ to take this into account. Steps 1 to 5 are repeated to obtain B analogs of the imputed total estimator. In step 6, the bootstrap variance of the imputed total is obtained using the standard bootstrap formulae.

Procedure 3 Variance of the imputed total estimator using the extended BWO of Sitter (1992a).

Step 1: Consider $n'_h = n_h - (1 - f_h)$ and $k_h = \frac{N_h}{n_h}\left(1 - \frac{1-f_h}{n_h}\right)$. If at least one of k_h or n'_h is not integer, apply a randomization (see Sitter 1992a, Section 3.1).

Step 2: Create the h-th stratum of the pseudopopulation by replicating k_h times the n_h units in S_h.

Step 3: Select a SRSWOR of size n'_h without replacement from the h-th stratum of the pseudopopulation created in step 2 to obtain a sample $S^*_h = \{hi : i = 1, \ldots, n'_h\}$.

Step 4: Repeat steps 1 to 3 independently for each stratum h to obtain a bootstrap sample $\mathbf{S}^* = \{S^*_1, \ldots, S^*_H\} = \{hi : h = 1, \ldots, H; i = 1, \ldots, n'_h\}$ of size $n^* = \sum_{h=1}^{H} n'_h$.

Step 5: Apply AM imputation to impute the bootstrap sample $\mathbf{S}^*$ and obtain the bootstrap analog of the imputed total estimator $\widehat{Y}_I$ by

$$\widehat{Y}_I^{(b)} = \frac{N}{N^*} \sum_{h=1}^{H} \sum_{hi \in S^*_h} \frac{\tilde{y}^{(*)}_{hi}}{n'_h/(n_h k_h)},$$

where $N^* = \sum_{h=1}^{H} n_h k_h$ and $\tilde{y}^{(*)}_{hi}$ is the value of the variable of interest of unit hi if this one is observed and the imputed value otherwise.

Step 6: Repeat steps 1 to 5 a large number of times B to obtain $\widehat{Y}_I^{(1)}, \ldots, \widehat{Y}_I^{(B)}$ where $\widehat{Y}_I^{(b)}$ is the analog of $\widehat{Y}_I$ for the b-th bootstrap sample.

Step 7: Obtain the bootstrap variance of $\widehat{Y}_I$ by

$$\mathrm{V}_{boot}(\widehat{Y}_I) = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{Y}_I^{(b)} - \widehat{Y}_I^{(.)}\right)^2,$$

where $\widehat{Y}_I^{(.)}$ is the mean bootstrap analog of $\widehat{Y}_I$

$$\widehat{Y}_I^{(.)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{Y}_I^{(b)}.$$

6 Simulations

A numerical study was conducted to test the performance of the proposed imputation method. Simulated data and real data were considered. In Sects. 6.1 and 6.2 , the simulation settings for the simulated data and for the real data are respectively presented. Measures used to compare the new imputation method with existing imputation methods and to test the accuracy of the bootstrap procedures for the variance estimation are described in Sect. 6.3. Finally, the results of the simulations in each setting are displayed and commented in Sects. 6.4 and 6.5 respectively.

6.1 Setting 1: Simulated data

Populations of size $N = 10{,}000$ were considered. Four auxiliary variables x_1, x_2, x_3, and x_4 were generated. The values x_{i1}, x_{i2}, and x_{i3}, $i = 1, \ldots, N$, are independent draws from a Uniform[0, 1] random variable and x_{i4}, $i = 1, \ldots, N$, are independent draws of a gamma density with shape and scale parameters, respec-

tively, 3 and 1/6 that were mapped into the [0, 1] interval via the transformation $x_{i4} \to (x_{i4} - \min(x_4)) / (\max(x_4) - \min(x4))$.

Five populations were then generated as follows:

$$
\begin{aligned}
y_i^{(1)} &= 1 + 5x_{i1} + x_{i2} + x_{i3} + x_{i4} + 10\varepsilon_i, \\
y_i^{(2)} &= 2 + \cos(\pi x_{i1} + \pi) + \sin(4\pi x_{i2}) + \exp(-(x_{i3} - 0.5)^2) + (x_{i4} - 0.5)^2 + 6\varepsilon_i, \\
y_i^{(3)} &= 1 + \cos(2\pi x_{i1}) + x_{i1}x_{i2} + x_{i3}^2 x_{i4} + 5\varepsilon_i, \\
y_i^{(4)} &= 2 + \cos(\pi(x_{i1} + x_{i2}))\sin(\pi(x_{i3} + x_{i4})) + \varepsilon_i, \\
y_i^{(5)} &= 1 + \varepsilon_i,
\end{aligned}
$$

where $i = 1, \ldots, N$, and where ε_i are N independent draws of a normal random variable with mean 0 and standard deviation 0.1. In the first four populations, the variable of interest is linked to the auxiliary variables. In the first two populations the link is correctly specified by an AM, even a linear model in population 1. In populations 3 and 4 the AM is not a valid representation of the truth, while in the last population there is no link between the variable of interest and the auxiliary variables.

Two different sampling designs were used for the selection of samples: simple random sampling without replacement (SRSWOR) and stratified sampling . For simple random sampling, a sampling rate of $f = 0.2$ was considered. For stratified sampling, strata were created as follows. First, units were classified into two groups, depending whether their value x_{i1} is larger than the median of x_1 or not. In each group created, units were then subdivided into two other groups, depending on whether their value x_{i2} is larger than the median of x_2 in each group or not. The procedure was repeated for variables x_3 and x_4. This resulted in creating 16 strata of size 625 that are somewhat homogeneous with respect to the auxiliary variables. Then, SRSWOR was applied within strata with a sampling rate of $f = 0.2$ in each stratum. Note that in this setting all units have the same design weight under both SRSWOR and stratified sampling. Therefore, unweighted imputation based on (5) and weighted imputation based on (6) yield the same results.

The response probabilities were obtained from

$$
p_i = \frac{\exp(b_0 + b_1 x_{i1})}{1 + \exp(b_0 + b_1 x_{i1})},
$$

where $b_0 = -1$ and $b_1 = 5$. These values yield an overall mean response rate of approximately 75%.

One thousand simulations were then conducted as follow. For each simulation, a sample S was selected according to either SRSWOR or stratified sampling. For each sample S selected, a respondents set S_r and a nonrespondents set S_m were then created by generating a response indicator vector $(r_i | i \in S)^\top$, where $r_i, i \in S$, was generated from a Bernoulli distribution with parameter p_i. Then, for each set of respondents and of nonrespondents obtained, the missing y_i, $i \in S_m$, were replaced with imputed y_i^* using the five following imputation methods:

- **Regression imputation** Imputed values $y_i^*, i \in S_m$, are obtained by

$$y_i^* = \widehat{\beta}_0 + \sum_{j=1}^{q} \widehat{\beta}_j x_{ij},$$

where $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_q)^\top$ is defined by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{j \in S_r} d_j (1, \mathbf{x}_j)^\top (1, \mathbf{x}_j) \right)^{-1} \sum_{i \in S_r} d_i (1, \mathbf{x}_i)^\top y_i.$$

Regression imputation is based on imputation model model 3 with $f(x_{i1}, x_{i2}, \ldots, x_{iq}) = \beta_0 + \sum_{j=1}^{q} \beta_j x_{ij}$.

- **Mean imputation** The missing $y_i, i \in S_m$, are replaced by the respondents' mean value, that is the imputed values $y_i^*, i \in S_m$, are obtained by

$$y_i^* = \frac{1}{\sum_{j \in S_r} d_j} \sum_{k \in S_r} d_k y_k.$$

Mean imputation is a particular case of regression imputation where only a constant covariate is considered. It is based on imputation model 3 with $f(x_{i1}, x_{i2}, \ldots, x_{iq}) = \beta_0$. For both simple random sampling and stratified sampling, we applied mean imputation within strata. This means that in the case of simple random sampling, strata were used as imputation classes in order to use the auxiliary information when imputing.

- **Nearest neighbor imputation** The missing $y_i, i \in S_m$, are replaced by their respective nearest neighbor in the complete data. The proximity is quantified through the auxiliary variables. Imputed values $y_i^*, i \in S_m$, are obtained by

$$y_i^* = y_{j(i)} \quad \text{where} \quad d(\mathbf{x}_i, \mathbf{x}_{j(i)}) = \min_{j \in S | r_j = 1} d(\mathbf{x}_i, \mathbf{x}_j),$$

where $d(\cdot, \cdot)$ is the Euclidean distance.

- **Random forest imputation** Random forest (Breiman 2001) is a nonparametric algorithm for classification and regression. It generates at random several classification trees. In the context of regression, predictions are obtained by averaging the output of all trees. The missing values were imputed with the nonparametric imputation method using random forest of Stekhoven and Buehlmann (2012). Imputation was carried out using function `missForest` of `R` package `missForest` (Stekhoven 2013). Function `missForest` begins with an initial guess for the missing values. Then, it sorts the variables according to the amount of missing values starting with the lowest amount. In our case, variable y is last since it is the only one with missing values. The missing values are imputed by first fitting a random forest to the observed values $(y_i, \mathbf{x}_i), i \in S_r$; then imputing

the missing values $y_i, i \in S_m$ by applying the trained random forest to $\mathbf{x}_i, i \in S_m$. The procedure is repeated until a stopping criterion is met.

- **AM imputation** An AM was fitted using the complete data $(y_i, \mathbf{x}_i)$, $i \in S_r$, and imputed values y_i^*, $i \in S_m$, were obtained through predictions with this model, as explained in Sect. 5. Survey weights were considered in the smoothing spline estimator computation of each term, as in the PLS equation of expression (6). The model was fitted using function `gam` of `R` package `mgcv` (Wood 2014). Function `gam` uses $m = 2$ and thin plate splines basis by default. The model is fitted by penalized likelihood maximization and the smoothing parameter is selected by generalized cross validation.

The imputed total estimator $\widehat{Y}_I$ was computed for each method and each simulation. Note that all the considered imputation methods use auxiliary information when computing imputed values under both SRSWOR and stratified sampling, including mean imputation. Indeed, mean imputation is applied in each stratum separately, and the strata are created using auxiliary information.

Moreover, one thousand simulations were conducted to test the accuracy of the bootstrap procedures presented in Sect. 5.2 to estimate the variance of the total. SRSWOR and stratified sampling were considered. For each simulation, a sample S, a set of respondents S_r and of nonrespondents S_m were created as described above. The missing values were replaced with imputed values using AM imputation. The imputed total estimator $\widehat{Y}_I$ and its bootstrap variance $\mathrm{V}_{boot}(\widehat{Y}_I)$ were computed for each simulation. For the bootstrap variance under SRSWOR, Procedure 1 (MMB) was applied where, in step 1, a sample of size 400 was selected, that is $n'_h = f \cdot n_h = 400$, $h = 1$ and Procedure 2 (BWO) was applied where, in step 1, $k = 25$. For the bootstrap variance under stratified sampling, Procedure 1 (MMB) was applied where, in step 1, a sample of size 125 was selected in each stratum, that is $n'_h = f \cdot n_h = 125$ for each stratum h and Procedure 3 (extended BWO) was applied where a randomization was applied in step 1. Note that randomization was applied only in the latest case. More cases where randomization is applied and a discussion on how randomization may affect the results are shown in Sect. 6.2 and in the "Appendix".

6.2 Setting 2: Expenditure data

We consider the data from the 1992 family expenditure survey (FES), see Central Statistical Office (1993). The data is made available by the UK data archive at the University of Essex. To test our method, we considered that the households having a non-missing and larger than zero disposable income (disposable income and self-supply and in kind) of the 1992 FES form the population of interest. The size of this population is $N = 7409$. The variable disposable income is highly skewed to the right (skewness is 3.15) and has many outliers. Important asymmetry and outliers may impact negatively the performance of all competing imputation methods and yield unstable imputed total estimators. For this reason, the variable disposable income was modified as follows. First, it was divided by its mean value. Then, the natural logarithm of the obtained value plus one was computed. One was added before computing the logarithm to avoid negative values. We suppose that the aim of the survey is to estimate

the population total of the modified disposable income. The population was stratified into 12 regions and simple random sampling with a sampling rate of $f = 0.2$ was applied within each region (stratum). The sample size was randomly rounded for 8 strata for which this sampling rate led to a non-integer sample size. For each sampled household, we supposed that the following characteristics were observed:

x_{i1}: number of adults in household i,
x_{i2}: number of children in household i,
x_{i3}: number of persons economically active in household i,
x_{i4}: age of the head of household i,
x_{i5}: age of the chief economic supporter of household i.

Such variables could for instance come from a register. Figure 1 shows the resulting data. The figure shows that none of the five characteristics listed above has a linear relationship with modified disposable income. It was supposed that the willingness of a household to respond depends on the number of adults in this household and that the households respond independently from each other. Hence, the response probabilities were obtained from

$$p_i = \frac{\exp(b_0 + b_1 x_{i1})}{1 + \exp(b_0 + b_1 x_{i1})},$$

where $b_0 = -1$ and $b_1 = 1$ which yields an overall mean response rate of approximately 70%. Then, for each sampled household, a response indicator was generated from a Bernoulli distribution with parameter p_i. The modified disposable income was then recorded for respondents and erased for nonrespondents. One thousand simulations were conducted. The same imputation methods as in Sect. 6.1 were considered.

Moreover, one thousand simulations were conducted to test the accuracy of the bootstrap procedures presented in Sect. 5.2 to estimate the variance of the total. For each simulation, a sample and a set of respondents and of nonrespondents were created as described above. The missing values were replaced with imputed values using AM imputation. The imputed total estimator $\widehat{Y}_I$ and its bootstrap variance $\mathrm{V}_{boot}(\widehat{Y}_I)$ were computed for each simulation. For the bootstrap variance, Procedure 1 (MMB) was applied with $B = 100$ bootstrap replicates. In step 1, we set $n'_h = \lceil f \cdot n_h \rceil$ where $\lceil \cdot \rceil$ is the ceiling function. In step 2, a randomization was applied to round the non-integer k_h (see Sitter 1992b, Section 1.6).

6.3 Measures of comparison

For each simulation and each imputation method of both settings, the population total for the variable of interest was estimated through the imputed estimator of expression (2). To compare the performance of the methods, four comparison measures were recorded. First, to quantify the accuracy of imputed values, the Monte Carlo mean relative prediction error was computed, which is defined as

$$\mathrm{MRPE} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{1}{n_m^{(\ell)}} \sum_{i \in S_m^{(\ell)}} \left| \frac{y_i^{*(\ell)} - y_i}{y_i} \right|,$$

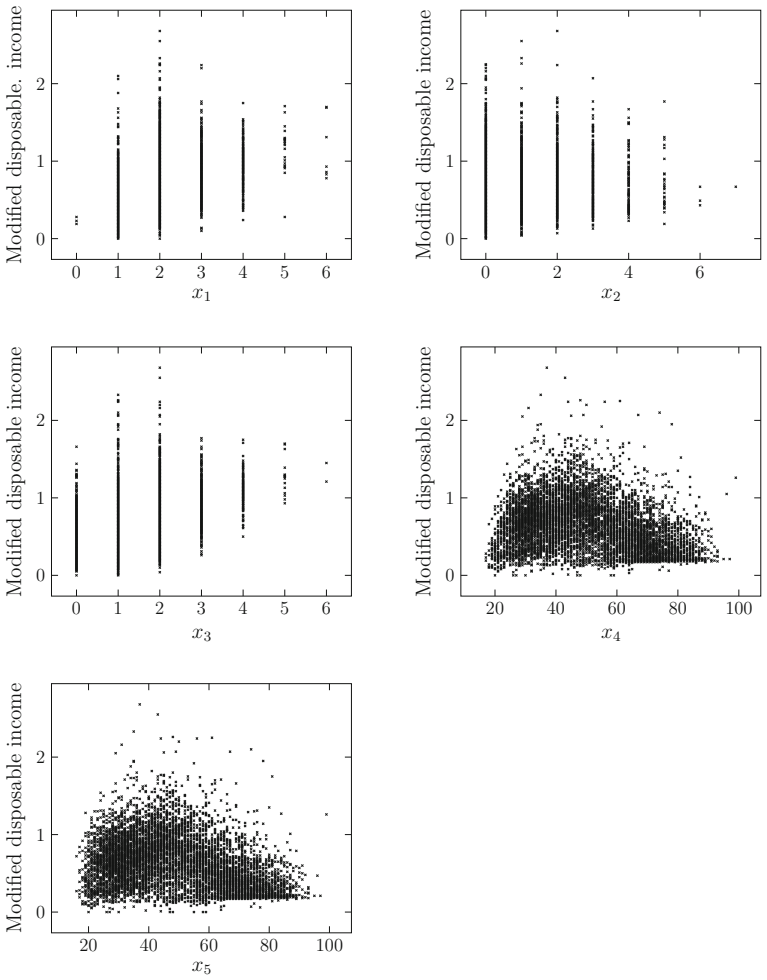

Fig. 1 Modified disposable income and number of adults in household (x_1), number of children in household (x_2), number of persons economically active in household (x_3), age of the head of household (x_4), and age of the chief economic supporter of household (x_5) in FES data

where $S_m^{(\ell)}$ is the nonrespondents set obtained at the ℓ-th simulation, $n_m^{(\ell)}$ is the size of $S_m^{(\ell)}$, $y_i^{*(\ell)}$ is the imputed value obtained for $i \in S_m^{(\ell)}$ at the ℓ-th simulation, and L represents the number of simulations. Then, for each imputation method, the performance of the imputed estimator of expression (2) was studied through three comparison measures, namely

- the Monte Carlo relative bias (RB) defined as

$$\mathrm{RB} = \frac{B}{Y},$$

where $\mathrm{B} = \widehat{Y}_I^{(\cdot)} - Y$, $\widehat{Y}_I^{(\cdot)}$ represents the mean imputed estimator over the L simulations

$$\widehat{Y}_I^{(\cdot)} = \frac{1}{L}\sum_{\ell=1}^{L} \widehat{Y}_I^{(\ell)},$$

and $\widehat{Y}_I^{(\ell)}$ is the imputed estimator $\widehat{Y}_I$ obtained at the ℓ-th simulation,

- the Monte Carlo relative root variance (or relative standard deviation) defined as

$$\mathrm{RRVAR} = \frac{(\mathrm{VAR})^{1/2}}{Y},$$

where

$$\mathrm{VAR} = \frac{1}{L-1}\sum_{\ell=1}^{L}\left(\widehat{Y}_I^{(\ell)} - \widehat{Y}_I^{(\cdot)}\right)^2,$$

- the Monte Carlo relative root mean square error defined as

$$\mathrm{RRMSE} = \frac{\left(\mathrm{B}^2 + \mathrm{VAR}\right)^{1/2}}{Y}.$$

For AM imputation, the following measures were computed to test the accuracy of the bootstrap variance estimator:

- The Monte Carlo variance of the total estimator:

$$\mathrm{VAR} = \frac{1}{L-1}\sum_{\ell=1}^{L}\left(\widehat{Y}_I^{(\ell)} - \widehat{Y}_I^{(\cdot)}\right)^2,$$

- The Monte Carlo expectation of the bootstrap variance estimator:

$$\mathrm{VAR}_{boot} = \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{V}_{boot}^{(\ell)}(\widehat{Y}_I),$$

where $\mathrm{V}_{boot}^{(\ell)}(\widehat{Y}_I)$ is the bootstrap variance $\mathrm{V}_{boot}(\widehat{Y}_I)$ obtained at the ℓ-th simulation,

- The coverage rate CR: the proportion of times the true total Y falls into the 95% confidence interval

$$\widehat{Y}_I \pm 1.96\sqrt{\mathrm{V}_{boot}(\widehat{Y}_I)}.$$

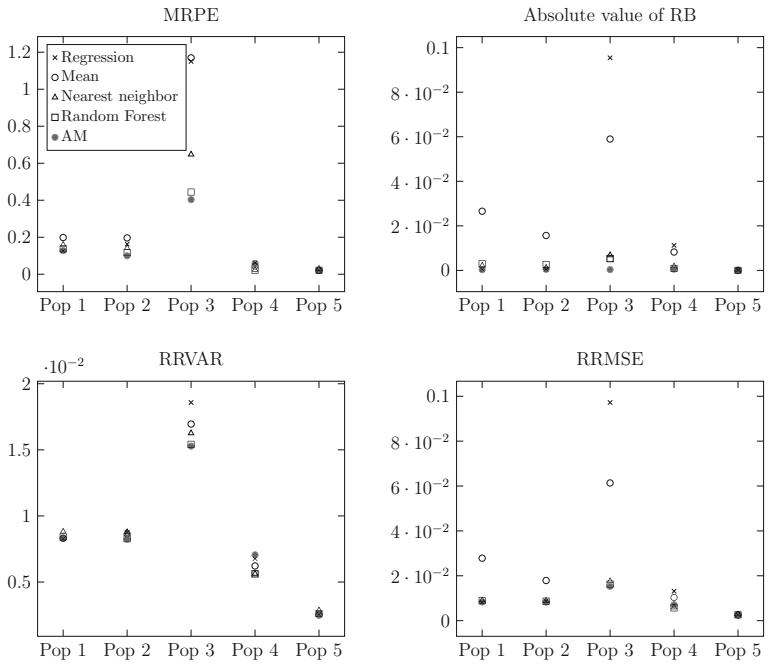

Fig. 2 Comparison measures of five imputation methods in five populations under SRSWOR

6.4 Results of setting 1

Figures 2, 3, and Table 2 display the results of setting 1. Table 1 reports the average ranks over the populations of each imputation method for each measure of comparison. The absolute value of RB was considered.

We first comment the results shown in Figs. 2 and 3. When functional dependence between the variable of interest and the auxiliary variables is additive (populations 1 and 2), AM imputation provides the best results. If, moreover, this functional dependence is linear (population 1), regression imputation performs as well as AM imputation. When there is no dependence between the variable of interest and the auxiliary variables (population 5), all five methods perform fairly similarly. Because the functional dependence between the variable of interest and the auxiliary variables is not additive in populations 3 and 4, the results for these two populations allow us to study the performance of AM imputation under model misspecification. We can see that AM imputation still performs the best in population 3. The reason for the good performance of AM imputation in this population is that, even though the functional dependence is not additive, it can be well approximated by an additive function. In population 4, the situation is less obvious and it is difficult to rank the imputation methods. In order to produce a global index of performance we ranked the imputing methods for each population and each performance criterion. The results, reported in Table 1 show that, globally, AM imputation performs better than the other imputation methods considered.

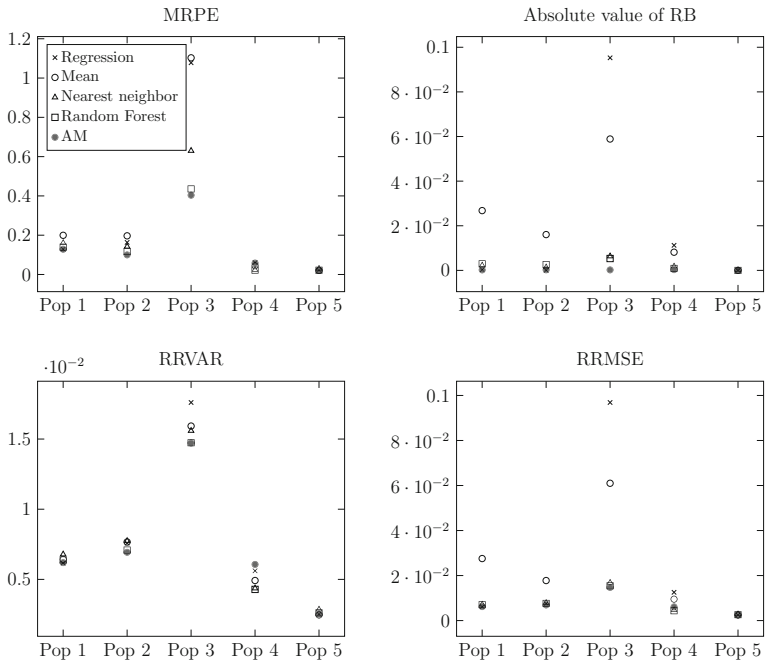

Fig. 3 Comparison measures of five imputation methods in five populations under stratified sampling

The performance of the bootstrap-based estimators of the variance is assessed in Table 2. Whether the functional dependence between the variable of interest and the auxiliary variables is additive (populations 1 and 2) or not (populations 3, 4, 5), the bootstrap variance is generally very close to the variance obtained by simulation and leads to very good coverage rates across all five populations considered. We see, however that BWO variance estimate tends to overestimate the variance when randomization is applied (here under stratified sampling). This overestimation is more pronounced when the functional dependence between the variable of interest and the auxiliary variables is additive and strong (populations 1 and 2). This phenomenon is discussed in the "Appendix".

6.5 Results of setting 2

Tables 3 and 4 display the results of our analysis performed under setting 2. The numbers in brackets in Table 3 report the rank of each imputation method for each measure of comparison. We can see that AM imputation outperforms the competing imputation methods. With this data, the bootstrap variance yields a coverage rate of 94% that is close to the theoretically stated value of 95%.

As we can see from the results of both settings, AM imputation performs the best overall, closely followed by random forest. This is not surprising since random forest is also nonparametric. Two advantage of random forest over our imputation method are: (1) it can handle mixed-type data and (2) auxiliary variables can have missing

Table 1 Average ranks over five populations of each imputation method for each measure of comparison (in absolute value)

Imputation method	MRPE	RB	RRVAR	RRMSE
Simple random sampling (SRSWOR)				
Regression	3.0	3.6	3.2	3.2
Mean	4.2	4.2	2.6	3.8
Nearest neighbor	3.4	2.8	4.0	3.6
Random forest	2.4	2.6	2.6	2.4
AM	2.0	1.8	2.6	2.0
Stratified sampling				
Regression	3.0	3.6	3.0	3.0
Mean	4.2	4.2	3.2	3.8
Nearest neighbor	3.4	3.2	4.0	3.6
Random forest	2.4	2.8	2.4	2.6
AM	2.0	1.2	2.4	2.0

Table 2 Monte Carlo variance of the total, Monte carlo expectation of the bootstrap variance and coverage rate associated with AM imputation for two different sampling designs and five populations

		MMB		BWO	
	$\text{VAR} \times 10^2$	$\text{VAR}_{boot} \times 10^2$	CR	$\text{VAR}_{boot} \times 10^2$	CR
Simple random sampling (SRSWOR)					
Population 1	157	152	0.94	151	0.94
Population 2	669	611	0.94	607	0.94
Population 3	396	378	0.94	381	0.94
Population 4	156	137	0.92	137	0.92
Population 5	7	6	0.94	6	0.94
Stratified sampling					
Population 1	896	843	0.94	1577	0.99
Population 2	495	447	0.93	636	0.97
Population 3	374	361	0.95	424	0.95
Population 4	110	106	0.93	148	0.95
Population 5	7	6	0.93	7	0.93

Table 3 Comparison measures for five imputation methods for FES data

Imputation method	$\text{MRPE} \times 10^1$	$\text{RB} \times 10^{-2}$	$\text{RRVAR} \times 10^{-2}$	$\text{RRMSE} \times 10^{-2}$
Regression	3.28(4)	0.79(3)	1.41(2)	1.61(3)
Mean	4.51(5)	5.57(5)	1.47(4)	5.76(5)
Nearest neighbor	3.19(3)	0.85(4)	1.54(5)	1.76(4)
Random forest	2.97(2)	0.22(2)	1.42(3)	1.43(2)
AM	2.90(1)	0.08(1)	1.39(1)	1.40(1)

Table 4 Monte Carlo variance of the total, Monte carlo expectation of the bootstrap variance and coverage rate associated with AM imputation for FES data

VAR	VAR_{boot}	CR
4394.03	4031.77	0.94

values. Two advantages of our method are: (1) it is fast and (2) it allows us to take design weights into account in the imputation model.

7 Conclusion

We propose new imputation method for nonresponse in surveys based on spline smoothing within the additive model paradigm. The simulations indicate that the new method is very flexible and can capture a large spectrum of functional dependencies between the variable of interest and the auxiliary variables. Since the model requires only weak assumptions, it is less susceptible to model misspecification than other models such as parametric ones. Most importantly, the AM formulation makes it possible to consider several auxiliary variables in the imputation process without running into the curse of dimensionality phenomenon. Bootstrap procedures to estimate the variance of the total under SRSWOR and stratified sampling was suggested.

Through a simulation study, the new imputation method was confirmed to perform well in many different situations. AM imputation performs better than the other imputation methods considered when the functional dependence between the variable of interest and the auxiliary variables is additive or when this dependence can be well approximated by an additive function. When this dependence is not well approximated by an additive function or when there is no dependence between the variable of interest and the auxiliary variables, AM imputation shows a performance similar to that of the other imputation methods considered. In most of the cases studied, the proposed bootstrap-based variance estimates were close to the true Monte Carlo variance and produced very good coverage rates. We explain why the proposed bootstrap-based variance estimates may overestimate the variance of the total estimator in some cases.

Future work include extending the current method to situations in which the samples are dependent and improving the computational speed of the variance via parallel processing.

Acknowledgements The authors thank Yves Tillé for his constructive suggestions. This research was supported by the Swiss National Science Foundation and the Natural Science and Engineering Research Council of Canada.

Appendix: Bootstrap variance when a randomization is applied

We repeated the simulations for the bootstrap variance of Sect. 6.1 with sampling fraction $f = 0.3$ in order to study the impact of randomization on the quality of variance estimates. For the bootstrap variance under SRSWOR, Procedure 1 (MMB)

Table 5 Monte Carlo variance of the total, Monte carlo expectation of the bootstrap variance and coverage rate associated with AM imputation for two different sampling designs and five populations

		MMB		BWO	
	$\text{VAR} \times 10^2$	$\text{VAR}_{boot} \times 10^2$	CR	$\text{VAR}_{boot} \times 10^2$	CR
Simple random sampling (SRSWOR)					
Population 1	914	873	0.95	877	0.95
Population 2	390	356	0.93	357	0.94
Population 3	244	222	0.93	220	0.93
Population 4	92	80	0.92	79	0.93
Population 5	4	4	0.93	4	0.93
Stratified sampling					
Population 1	525	2643	1.00	918	0.99
Population 2	282	790	1.00	370	0.97
Population 3	219	259	0.97	243	0.96
Population 4	79	163	1.00	83	0.95
Population 5	4	4	0.92	4	0.92

was applied where, in step 1, a sample of size 900 was selected, that is $n'_h = f \cdot n_h = 900$, $h = 1$ and a randomization was applied in step 2, and Procedure 3 (extended BWO) was applied (k was non-integer) where a randomization was applied in step 1. For the bootstrap variance under stratified sampling, Procedure 1 (MMB) was applied where, in step 1, a sample of size 187 was selected in each stratum, that is $n'_h = \lfloor f \cdot n_h \rfloor = 187$, where $\lfloor \cdot \rfloor$ is the floor function, for each stratum h and a randomization was applied in step 2, and Procedure 3 (extended BWO) was applied where a randomization was applied in step 1. Note that randomization was applied in all four cases.

Table 5 shows the result. Under SRS, whether the functional dependence between the variable of interest and the auxiliary variables is additive (populations 1 and 2) or not (populations 3, 4, 5), the bootstrap variance is close to the variance obtained by simulation and it leads to very good coverage rates (between 92% and 94%) across all five populations considered. Under stratified sampling, the bootstrap variance is greater than the variance obtained by simulations in four out of the five populations considered. This difference is greater when the functional dependence between the variable of interest and the auxiliary variables is additive and strong (populations 1 and 2). We explain this phenomenon in what follows.

When a randomization is applied to round the non-integer k_h and/or n'_h as it is the case here, the bootstrap variance contains two parts: the variance due to the randomization and the variance of the total estimator. When there is a strong additive functional dependence between the variable of interest and the auxiliary variables, the variance of the total estimator is small. An important portion of the bootstrap variance is due to randomization and the bootstrap variance overestimates the variance of the total. As the additive functional dependence between the variable of interest and the auxiliary variables weakens, the variance of the total estimator increases and the portion of the bootstrap variance due to randomization decreases. The bootstrap variance gets closer

to the variance of the total. When stratified sampling is applied, the portion of the variance due to randomization may be particularly important because randomization is applied within each stratum. This explains the difference between the bootstrap variance and the variance obtained by simulations under stratified sampling in Table 5. The simulations run on the real data of Sect. 6.2 confirm this explanation. In this setting, there is a moderate additive functional dependence between the variable of interest and the auxiliary variables. Stratified sampling was used and the randomization procedure was applied to round the non-integer quantities. The obtained bootstrap variance is close to the variance obtained by simulations and yields a coverage rate of 94%.

As shown by these results, randomization affects the quality of the variance estimates. We refer the reader to Andreis et al. (2018) about weights rounding problems in resampling. We repeated the simulation in this section and rounded the non-integer k_h and n'_h to the nearest integer instead of applying randomization. This yields very similar results.

References

Andreis F, Conti PL, Mecatti F (2018) On the role of weights rounding in applications of resampling based on pseudopopulations. Stat Neerl

Andridge RR, Little RJA (2010) A review of dot deck imputation for survey non-response. Int Stat Rev 78:40–64

Bellman R (1961) Adaptive control processes: a guided tour. Princeton University Press, Princeton

Berg E, Kim J-K, Skinner C (2016) Imputation under informative sampling. J Surv Stat Methodol 4(4):436–462

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Central Statistical Office (1993) Family expenditure survey, 1992 [computer file]. Technical report, Colchester, Essex: UK Data Archive [distributor]. SN: 3064. https://doi.org/10.5255/UKDA-SN-3064-1

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

Da Silva DN, Opsomer JD (2006) A kernel smoothing method of adjusting for unit non-response in sample surveys. Can J Stat 34(4):563–579

Da Silva DN, Opsomer JD (2009) Nonparametric propensity weighting for survey nonresponse through local polynomial regression. Surv Methodol 35(2):165–176

Eubank RL (1999) Nonparametric regression and spline smoothing, 2nd edn. Marcel Dekker, New York

Giommi A (1987) Nonparametric methods for estimating individual response probabilities. Surv Methodol 13(2):127–134

Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models. Chapman & Hall, Boca Raton

Gross ST (1980) Mean estimation in sample surveys. In: Proceedings of the survey research methods section. American Statistical Association, pp 181–184

Hastie TJ, Tibshirani RJ (1986) Generalized additive models. Stat Sci 1(3):297–318

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman & Hall, Boca Raton

Haziza D (2009) Imputation and inference in the presence of missing data. In: Rao C (ed) Handbook of statistics, volume 29 of handbook of statistics. Elsevier, Amsterdam, pp 215–246

Haziza D, Rao JNK (2005) Inference for domain means and totals under imputation for missing data. Can J Stat 33:149–161

Lee TCM (2003) Smoothing parameter selection for smoothing splines: a simulation study. Comput Stat Data Anal 42(1):139–148

Mashreghi Z, Léger C, Haziza D (2014) Bootstrap methods for imputed data from regression, ratio and hot-deck imputation. Can J Stat 42(1):142–167

Ning J, Cheng P (2012) A comparison study of nonparametric imputation methods. Stat Comput 22:273–285

Niyonsenga T (1994) Nonparametric estimation of response probabilities in sampling theory. Surv Methodol 20(2):177–184
Niyonsenga T (1997) Response probability estimation. J Stat Plan Inference 59:111–126
Qin J, Leung D, Shao J (2002) Estimation with survey data under nonignorable nonresponse or informative sampling. J Am Stat Assoc 97(457):193–200
Rubin DB (1976) Inference and missing data. Biometrika 63:581–592
Särndal C-E (1992) Methods for estimating the precision of survey estimates when imputation has been used. Surv Methodol 18(2):241–252
Shao J, Sitter RR (1996) Bootstrap for imputed survey data. J Am Stat Assoc 91:1278–1288
Sitter RR (1992a) Comparing three bootstrap methods for survey data. Can J Stat 20:135–154
Sitter RR (1992b) A resampling procedure for complex survey data. J Am Stat Assoc 87(416):755–765
Stekhoven DJ (2013) missForest: nonparametric missing value imputation using random forest. R package version 1:4
Stekhoven D, Buehlmann P (2012) Missforest—nonparametric missing value imputation for mixed-type data. Bioinformatics 28(1):112–118
Stones CJ (1985) Additive regression and other nonparametric models. Ann Stat 13(2):689–705
Wang Y (2011) Smoothing splines: methods and applications. Chapman & Hall, Boca Raton
Wood S (2003) Thin plate regression splines. J R Stat Soc Ser B (Stat Methodol) 65(1):95–114
Wood S (2008) Fast stable direct fitting and smoothness selection for generalized additive models. J R Stat Soc Ser B (Stat Methodol) 70(3):495–518
Wood S (2014) mgcv: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation. R package version 1.7-28. http://CRAN.R-project.org/package=mgcv
Zhang G, Christensen F, Zheng W (2013) Nonparametric regression estimators in complex surveys. J Stat Comput Simul 85(5):1026–1034