Statistical Inference in the Presence of Imputed Survey Data Through Regression Trees and Random Forests

Mehdi Dagdoug[(a)], Camelia Goga[(b)] and David Haziza[(c)*]

(a) McGill University, Department of Mathematics and Statistics, Montréal, CANADA

(b) Université de Franche-Comté, LMB, Besançon, FRANCE

(c) University of Ottawa, Department of Mathematics and Statistics, Ottawa, CANADA

Abstract

Item nonresponse in surveys is usually handled through some form of imputation. In recent years, imputation through machine learning procedures has attracted a lot of attention in national statistical offices. However, little is known about the theoretical properties of the resulting point estimators in a survey setting. In this paper, we study regression trees and random forests that provide flexible tools for obtaining imputed values. In a high-dimensional framework allowing the number of predictors to diverge, we lay out a set of conditions for establishing the mean square consistency of regression trees and random forests imputed estimators of a finite population mean. We propose a novel variance estimator based on a K-fold cross-validation procedure. The proposed point and variance estimation are assessed through a simulation study in terms of bias, efficiency, and coverage rate of normal-based confidence intervals. Finally, the choice of hyper-parameters involved in random forest algorithms is investigated through theoretical and empirical work.

Key words: Imputation; Item nonresponse; Missing data; Regression trees; Random forest; Variance estimation; Cross-validation.

*Address for correspondence: David Haziza, Department of mathematics and statistics, *University of Ottawa, Ottawa, Canada. Email: dhaziza@uottawa.ca*

1 Introduction

Since the seminal paper of Breiman (2001), random forests have been used in a variety of applications, including medicine (Fraiwan et al., 2012), time series analysis (Kane et al., 2014), agriculture (Grimm et al., 2008), missing data (Stekhoven and Buhlmann, 2011), genomics (Qi, 2012) and pattern recognition (Rogez et al., 2008), among others. Random forests constitute a class of ensemble models based on a large collection of B trees. Predictions through random forests are obtained by averaging the predictions obtained from each of the B trees of a forest.

Several empirical studies have shown that random forests compare favorably to other nonparametric methods (Hamza and Larocque, 2005; Díaz-Uriarte and de Andrés, 2006; Dagdoug et al., 2023a). Moreover, unlike several nonparametric statistical procedures (e.g., kernel regression, k-nearest neighbors, splines), random forests perform relatively well in high-dimensional and sparse settings; see, e.g., (Biau, 2012; Scornet et al., 2015; Klusowski and Tian, 2024).

Establishing the theoretical properties of random forests is challenging. In a non survey sampling setup, some important theoretical developments have been made by Scornet et al. (2015), who studied the mean square consistency of random forest predictors using results from Nobel (1996), assuming a fixed number of predictors. More recently, new high-dimensional results for random forests have been obtained, e.g., Klusowski and Tian (2024); Chi et al. (2022). Random forests have also been studied through the theory of U-statistics, e.g., Mentch and Hooker (2016); Zhou et al. (2019); Xu et al. (2024).

In finite population sampling, regression trees and random forests have also been applied in a variety of setups: (i) Model-assisted estimation (Tipton et al., 2013; De Moliner and Goga, 2018; McConville and Toth, 2019; Dagdoug et al., 2023b); (ii) Small area estimation (Krennmair and Schmid, 2022; Michal et al., 2023); (iii) The treatment of unit nonresponse (Phipps and Toth, 2012; Earp et al., 2018; Lohr et al., 2015); (iv) Design-based prediction (Toth and Eltinge, 2011; Nalenz et al., 2024).

This paper aims to provide a theoretical investigation of the properties of imputed estimators in surveys based on regression trees and random forests. The problem of missing data in surveys is ubiquitous. Estimators of population means based on complete cases tend to exhibit significant biases when the proportion of missing data is appreciable and

the behavior of the responding units differs from that of the nonresponding units. In this paper, we consider the problem of item nonresponse, a term used to describe the absence of information on some, but not all, survey variables for a sample unit. The missing values are imputed using a plausible value constructed based on auxiliary variables available for respondents and nonrespondents. Imputation procedures share a common objective: reduce the potential nonresponse bias to the best possible extent. Every imputation procedure relies on some implicit or explicit assumptions about the distribution of the survey variable requiring imputation. This set of assumptions is called an imputation model. The reader is referred to Haziza (2009) and Chen and Haziza (2019) for comprehensive discussions of imputation procedures in survey sampling. Tree-based methods such as random forests may prove useful for obtaining a set of imputed values. Because they are nonparametric, they tend to be robust against model misspecification. Also, with the emergence of large data sets in National Statistical Offices (NSO), random forests have attracted much attention in recent years and they are currently being scrutinized as an alternative to traditional imputation procedures. Recently, Dagdoug et al. (2023a) conducted an extensive simulation study to assess the performance of several machine-learning imputation procedures in terms of bias and efficiency in a wide variety of settings, including tree-based methods such as regression trees and random forests (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), Bayesian additive regression trees (Chipman et al., 2010) and Cubist (Quinlan, 1993)). The results of Dagdoug et al. (2023a) confirm the good performance of random forests.

Variance estimation is an important issue, as NSOs publish point estimates as well as corresponding estimated coefficients of variation, defined as the ratio of the estimated standard error of the estimate to the point estimate. Treating imputed values as observed values and applying a complete data variance estimation procedure will typically result in serious underestimation of the true variance of the imputed estimators. The resulting estimated coefficients of variation will thus be too small and the confidence intervals too narrow. As a result, inferences may be misleading. This has led researchers to develop a variety of variance estimators that account for sampling and nonresponse; see Haziza and Vallée (2020) for a comprehensive overview of approaches for estimating the variance of point estimators based on observed and imputed data. Because imputation procedures based on machine learning procedures may suffer from the problem of overfitting, applying customary variance estimators based on a first-order Taylor expansion may lead to appreciable underestimation

of the variance of point estimators. With complete data, this issue has been discussed by Opsomer and Miller (2005) and Dagdoug et al. (2023b) in the case of model-assisted estimation based on local polynomials and random forests, respectively. In a model-assisted framework, Dagdoug et al. (2023b) suggested a novel variance estimator based on a K-fold cross-validation procedure that prevents overfitting. We extend this K-fold procedure to estimate the variance of imputed estimators based on regression trees and random forests. The proposed variance estimator prevents from overfitting and is shown to perform well in a wide variety of settings.

The outline of the article is as follows. In Section 2, we define the framework and introduce the notation. In Section 3, we present the regression trees based on the CART algorithm (Breiman et al., 1984) and two random forest algorithms: the uniform random forest algorithm (Biau et al., 2008), and the algorithm of Breiman (Breiman, 2001). In Section 4, we establish some asymptotic properties of imputed estimators based on regression trees and random forests. We show that these imputed estimators are mean square consistent, even in a high-dimensional setting. In Section 5, using the reverse approach of Shao and Steel (1999), we describe two variance estimators: the first is obtained through a first-order Taylor expansion and is based on sample residuals, whereas the second variance estimator relies on residuals obtained by a K-fold cross-validation procedure. In Section 6, we present the results of a simulation study evaluating the performances of the proposed point and variance estimators in terms of bias, efficiency, and coverage rate of normal-based confidence intervals. The choice of some important hyper-parameters is discussed in Section 8. Finally, we make some concluding remarks in Section 8. Additional simulation studies as well as all proofs and further technical details are provided in the Appendix.

2 The setup

Consider a finite population $U = \{1, 2, ..., N\}$ of known size N. We are interested in estimating the finite population mean

$$\mu := \frac{1}{N} \sum_{k \in U} y_k,$$

of a survey variable Y, where y_k denotes the y-value for the k-th unit, $k \in U$. We select a sample S, of size n, according to a sampling design $\mathcal{P}$ with first-order inclusion probabilities $\{\pi_k\}_{k\in U}$ and second-order inclusion probabilities $\{\pi_{k\ell}\}_{k\neq\ell\in U}$ defined as $\pi_k := \mathbb{P}(k \in S)$ and $\pi_{k\ell} := \mathbb{P}(k, \ell \in S)$ for all $k, \ell \in U$. The sample S is completely characterized by the vector of sample selection indicators $\boldsymbol{I} = (I_1, \ldots, I_k, \ldots, I_N)^\top$, where $I_k := 1$ if $k \in S$ and $I_k := 0$, otherwise.

In the ideal case of complete response, provided that $\pi_k > 0$ for all $k \in U$, the Horvitz-Thompson estimator of μ defined by

$$\widehat{\mu}_\pi := \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k}. \tag{1}$$

is design-unbiased. In practice, the Y variable may be prone to missingness. Let $\boldsymbol{r} := (r_1, \ldots, r_k, \ldots, r_N)^\top$ denote the vector of response indicators such that $r_k := 1$ if y_k is observed, and $r_k := 0$ otherwise. Let $S_r := \{k \in S \ ; \ r_k = 1\}$ be the set of respondents to item Y, of size n_r, and $S_m := \{k \in S \ ; \ r_k = 0\}$ be the set of nonrespondents, of size n_m, such that $S_r \cup S_m = S$ and $n_r + n_m = n$. Let $\mathbf{x}_k := (x_{k1}, \ldots, x_{kp})^\top$ be the p-vector of measurements associated with p auxiliary variables $X_1, \ldots, X_p$ recorded for all $k \in S$, and let $\boldsymbol{X} := (\mathbf{x}_k^\top)_{k\in U}$ denote the corresponding population matrix. The observed data are given by

$$D_{imp} := \Big\{ (\mathbf{x}_k, y_k) \ ; \ k \in S_r \Big\} \bigcup \Big\{ \mathbf{x}_k \ ; \ k \in S_m \Big\}.$$

In this article, we restrict our attention to non-informative designs; see e.g., Pfeffermann and Sverchkov (2009). Assuming the design variables are available at the imputation stage, non-informativeness can be achieved by incorporating into the imputation model the subset of design variables associated with the variable Y requiring imputation. If the design information is unavailable at the imputation stage, the sampling weight can be included in the imputation model as a predictor; see Berg et al. (2016) for a discussion. The vectors $\left(r_k, y_k, \mathbf{x}_k^\top\right)_{k\in U}$ are assumed to be independent and identically distributed (i.i.d.). We further assume that the nonresponse model satisfies: (i) The positivity assumption, i.e., $\mathbb{P}\left(r_k = 1|\mathbf{x}_k\right) > 0$, almost surely; (ii) The Missing At Random (MAR) assumption (Rubin, 1976), i.e., $\mathbb{P}(r_k = 1|\ y_k, \mathbf{x}_k) = \mathbb{P}(r_k = 1|\ \mathbf{x}_k)$. The relationship between the survey variable

and the predictors may be described through the following imputation model:

$$\xi : \qquad y_k = m(\mathbf{x}_k) + \epsilon_k, \quad k \in S_r, \tag{2}$$

where $m(\mathbf{x}) := \mathbb{E}[Y|\mathbf{x}]$ denotes the unknown regression function and the errors $\{\epsilon_k\}_{k\in U}$ are i.i.d. random variables satisfying $\mathbb{E}[\epsilon_k|\mathbf{x}_k] = 0$ and $\mathbb{V}(\epsilon_k|\mathbf{x}_k) = \sigma^2 < \infty$.

Consider an estimator $\widehat{m} : \mathbb{R}^p \to \mathbb{R}$ of m fitted on $D_r := \{(\mathbf{x}_k, y_k) \ ; \ k \in S_r\}$. An imputed estimator $\widehat{\mu}_{\widehat{m}}$ of μ, based on the imputation procedure $\widehat{m}$, is given by

$$\widehat{\mu}_{\widehat{m}} := \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}(\mathbf{x}_k)}{\pi_k}\right), \tag{3}$$

where $\widehat{m}(\mathbf{x}_k)$ denotes the imputed value associated with $k \in S_m$.

3 Regression tree and random forest imputation

3.1 Regression tree imputation

Regression trees based on the CART algorithm (Breiman et al., 1984) are simple to implement and easily interpretable. With regression trees, the predictions are first obtained by partitioning the predictor space spanned by the predictors on D_r according to some criterion into disjoint regions called terminal nodes or leaves. Then, the prediction $\widehat{m}_{tree}$ at a point $\mathbf{x}$ belonging to a terminal node, denoted $A(\mathbf{x})$, is obtained by averaging the y-values recorded for the respondents in the same node:

$$\widehat{m}_{tree}(\mathbf{x}) := \frac{1}{N(\mathbf{x}, S_r)}\sum_{k\in S_r}\mathbb{1}_{\{\mathbf{x}_k\in A(\mathbf{x})\}}y_k, \tag{4}$$

where $N(\mathbf{x}, S_r) := \sum_{l\in S_r}\mathbb{1}_{\{\mathbf{x}_l\in A(\mathbf{x})\}}$ denotes the number of elements of S_r in the node $A(\mathbf{x})$ containing $\mathbf{x}$. More generally, for any subset $B \subseteq S$, we use the following notations

$$N(\mathbf{x}, B) := \sum_{\ell\in B}\mathbb{1}_{\{\mathbf{x}_\ell\in A(\mathbf{x})\}}, \qquad \widehat{N}(\mathbf{x}, B) := \sum_{\ell\in B}\frac{\mathbb{1}_{\{\mathbf{x}_\ell\in A(\mathbf{x})\}}}{\pi_l}, \tag{5}$$

to denote the unweighted and the weighted number of elements of B belonging to the node containing $\mathbf{x}$, respectively.

An imputed estimator of μ based on regression trees is thus given by

$$\widehat{\mu}_{tree} := \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}_{tree}(\mathbf{x}_k)}{\pi_k}\right). \tag{6}$$

The greedy CART algorithm (Breiman et al., 1984) is a popular method that recursively searches for the splitting variable and the splitting position (i.e., the coordinates on the predictor space where to split), leading to the largest possible reduction in the residual mean of squares before and after splitting. More specifically, let $\mathcal{C}_A$ be the set of all possible pairs (j,z) = (variable, position) in A and $A_L(j,z) = \{\mathbf{x}_k \in A; x_{kj} < z\}$, $A_R(j,z) = \{\mathbf{x}_k \in A; x_{kj} \geqslant z\}$. The best split (j^*, z^*) in a region A is defined as an element in $(j^*, z^*) \in \arg\min_{(j,z)\in\mathcal{C}_A}\{\sum_{k\in S_r:\mathbf{x}_k\in A_L(j,z)}(y_k-\bar{y}_{A_L})^2+\sum_{k\in S_r:\mathbf{x}_k\in A_R(j,z)}(y_k-\bar{y}_{A_R})^2\}$, where $\bar{y}_{A_L}$ (respectively $\bar{y}_{A_R}$) is the average of the y-values of units belonging to the node $A_L(j,z)$ (respectively $A_R(j,z)$). Splits are always performed in the middle of two points. The procedure continues until a stopping criterion is reached. Common stopping criteria include specifying the minimum number of elements, n_0, to be contained in each terminal node, or a maximal depth of the tree, K. For more details about trees and partitioning procedures, the reader is referred to Hastie et al. (2011) and Györfi et al. (2006). Since the CART algorithm uses the data $\{(\mathbf{x}_k, y_k)\}_{k\in S_r}$ to partition the predictor space, the resulting mutually exclusive regions depend on $\{(\mathbf{x}_k, y_k)\}_{k\in S_r}$. Alternative algorithms that do not make use of the survey variable Y to create the partition have been studied in the literature. This type of algorithm is said to have the X-property (Devroye et al., 2013).

Remark 3.1. *The estimator $\widehat{\mu}_{tree}$ can be expressed as*

$$\widehat{\mu}_{tree} = \frac{1}{N}\sum_{k\in S_r} w_k y_k,$$

with weights w_k given by

$$w_k = \frac{1}{\pi_k} + \frac{\widehat{N}(\mathbf{x}_k, S_m)}{N(\mathbf{x}_k, S_r)},$$

where $\hat{N}(\mathbf{x}_k, S_m)$, denotes the weighted number of units in S_m belonging to the same leaf as unit k, and $N(\mathbf{x}_k, S_r)$ denotes its unweighted version on S_r; see Equation (5). Since the partitioning algorithm usually uses the survey variable Y to make its splits, the weights $\{w_k\}_{k\in S_r}$ are dependent on Y. Also, in the case of an equal probability sampling design, the

tree imputed estimator $\widehat{\mu}_{tree}$ can be written in the so-called projection form:

$$\widehat{\mu}_{tree} = \sum_{k \in S} \widehat{m}_{tree}\left(\mathbf{x}_k\right).$$

This property not longer holds for unequal probability sampling design unless the predictions in each node are weighted by the inverse of the first-order inclusion probabilities.

3.2 Random forest imputation

Deep regression trees are simple to interpret and often exhibit a small model bias. However, they may suffer from a large variance as they tend to overfit the data. Breiman (1996) introduced the concept of bagged predictions, which consists of averaging predictions built on a large collection of regression trees fitted on different bootstrap samples from the original data. Random forests (Breiman, 2001) suggested adding even more diversity by considering at each split a subset of p_0 predictors selected randomly from the initial p predictors, as split candidates. The randomness induced by the bootstrap sampling procedure and the random selection of p_0 predictors at each split may be formalized by introducing a random variable Θ independent of the data and defined on some measurable space. The random forest predictions of $m(\mathbf{x})$ are thus a function of Θ; see e.g., Biau et al. (2008) for details.

Let $\widehat{m}_{tree}(\cdot, \Theta^{(b)})$ be the estimator of m based on the b-th randomized tree, $b = 1, \ldots, B$, where $\left\{\Theta^{(b)}\right\}_{b=1}^{B}$ is a set of i.i.d. random variables with distribution $\mathbb{P}_{\Theta}$. For simplicity, we write $\widehat{m}_{tree}^{(b)}$ for $\widehat{m}_{tree}(\cdot, \Theta^{(b)})$. The random forest prediction at $\mathbf{x}$ is defined as

$$\widehat{m}_{rf}^{(B)}(\mathbf{x}) := \frac{1}{B} \sum_{b=1}^{B} \widehat{m}_{tree}^{(b)}(\mathbf{x}). \tag{7}$$

The imputed y-values are given by $\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)$, $k \in S_m$. An imputed estimator of μ based on random forests with B trees is thus defined as

$$\widehat{\mu}_{rf}^{(B)} := \frac{1}{N} \left(\sum_{k \in S_r} \frac{y_k}{\pi_k} + \sum_{k \in S_m} \frac{\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)}{\pi_k} \right). \tag{8}$$

We adopt the following notation: The node of the b-th tree containing element $k \in S$ is denoted $A_b\left(\mathbf{x}_k\right)$. Also, for $k \in S$, we write $\psi_k^{(b)} = 1$ if unit k is selected in subsample $S_r\left(\Theta_b\right)$ of tree b, and $\psi_k^{(b)} = 0$ otherwise.

The forest imputed estimator $\widehat{\mu}_{rf}^{(B)}$ given by (8) can be expressed as the average of (randomized) tree imputed estimators,

$$\widehat{\mu}_{rf}^{(B)} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{tree}^{(b)}, \tag{9}$$

where $\widehat{\mu}_{tree}^{(b)}$ denotes the imputed estimator of μ given by (6). Therefore, many of the properties of tree imputed estimators are also shared by random forest imputed estimators. It is worth pointing out that the random forest imputed estimator $\widehat{\mu}_{rf}^{(B)}$ can be written as

$$\widehat{\mu}_{rf}^{(B)} = \frac{1}{N}\sum_{k\in S_r} w_k^{(B)} y_k, \qquad \text{where} \qquad w_k^{(B)} = \frac{1}{\pi_k} + \frac{1}{B}\sum_{b=1}^{B}\psi_k^{(b)}\frac{\widehat{N}_b\left(\mathbf{x}_k, S_m\right)}{N_b\left(\mathbf{x}_k, S_r\left(\Theta_b\right)\right)},$$

with N_b and $\widehat{N}_b$ defined as in (5) for the b-th tree of the forest. In the case of an equal probability sampling design, we can write $\widehat{\mu}_{rf}^{(B)}$ as

$$\widehat{\mu}_{rf}^{(B)} = \frac{1}{N}\left\{\sum_{k\in S}\widehat{m}_{rf}^{(B)}(\mathbf{x}_k) + \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S_r}\left(1-\psi_k^{(b)}\right)\frac{y_k - \widehat{m}_{tree}^{(b)}\left(\mathbf{x}_k\right)}{\pi_k}\right\}. \tag{10}$$

The quantity $(1-\psi_k^{(b)})(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k))$, in (10), is non zero for units $k \in S_r$ such that $\psi_k^{(b)} = 0$, i.e., for units $k \in S_r$ not selected on the b-th bootstrap sample. These units are referred to as *out-of-bag* observations since they are not used in the model fitting. The second term on the right-hand side of (10) serves as a bias-correction or debiasing term. A similar result has been obtained by Dagdoug et al. (2023b) in case of model-assisted random forest estimator. Notably, the form (10) holds even in the case of unequal probability sampling. However, in this case, the sum of residuals over $S_r\left(\Theta_b\right)$ may not be equal to zero unless the predictions from each tree are weighted by using the sampling weights.

In this article, we consider two random forest algorithms: (i) the uniform random forest algorithm; (ii) the random forests of Breiman. Uniform random forests are primarily studied in the literature because the partitions of the trees are independent of the observed data, thus making their theoretical analysis much simpler. However, because they do not use the data for building the partitions, they are of little practical interest. In practice, Breiman's original algorithm is typically used, but establishing its theoretical properties is more challenging.

Uniform random forests (Biau et al., 2008; Scornet, 2016) are based on a simple algorithm whereby the partition is created independently of the observed data. More precisely, given

a depth K, the partitioning algorithm splits each cell exactly K times, where each split is performed by choosing uniformly at random a variable among the initial p predictors $X_1, X_2, \ldots, X_p$ and a location along that coordinate uniformly at random.

As mentioned above, the randomness in Breiman's algorithm arises from constructing each of the B trees on different bootstrap samples and by considering as splitting candidates, at each split, p_0 predictors selected uniformly at random among the initial p predictors. We consider a slight modification of the algorithm that consists of selecting the B random samples without replacement (i.e., subsampling) of size a_v instead of sampling with replacement. This is common in the literature as it simplifies the theoretical derivations and is known to lead to efficient estimators; see, e.g., Bühlmann and Yu (2002) and the references therein for details. Finally, in Breiman's algorithm, splits are performed up to reaching one of the following two conditions: (i) there is only one respondent in the node considered; (ii) the maximal depth, K, is reached.

4 Theoretical results

In this section, we establish the mean square consistency of imputed estimators based on regression trees and random forests.

4.1 General assumptions

We consider the asymptotic framework of Isaki and Fuller (1982). Let $\{U_v\}_{v\in\mathbb{N}}$ denote a sequence of embedded finite populations of size $\{N_v\}_{v\in\mathbb{N}}$. In each finite population U_v, a sample S_v, of size n_v, is selected according to a sampling design $\mathcal{P}_v$ with inclusion probabilities $\pi_{k,v}$ and $\pi_{k\ell,v}$. While the finite populations are assumed to be embedded, we do not require this property to hold for the samples $\{S_v\}_{v\in\mathbb{N}}$. Imputation within sample S_v is performed using p_v predictors. This asymptotic framework implies that, as v goes to infinity, the population size N_v, the sample size n_v, as well as the number of predictors p_v increase to infinity. To improve readability, we use the subscript v only in the quantities U_v, N_v, n_v and p_v; quantities such as $\pi_{k,v}$ and $\pi_{k\ell,v}$ will simply be denoted by π_k and $\pi_{k\ell}$, respectively.
We start by describing a set of regularity conditions pertaining to the sampling design.

(H1) We assume that the sequence of sampling designs $\{\mathcal{P}_v\}_{v\in\mathbb{N}}$ is non-informative and that:

a) The sampling fraction is such that $\lim_{v\to\infty} \frac{n_v}{N_v} = \pi^* \in (0;1)$.

b) There exists positive constants λ and λ^* such that, for all $v \in \mathbb{N}$, $\min_{k\in U_v} \pi_k \geqslant \lambda > 0$, $\min_{k,\ell\in U_v} \pi_{k\ell} \geqslant \lambda^* > 0$.

c) The sample indicators covariances, $\Delta_{k\ell} := \pi_{k\ell} - \pi_k\pi_\ell$ for $k \neq \ell \in U_v$, are such that $\limsup_{v\to\infty} n_v \max_{k\neq\ell\in U_v} |\pi_{k\ell} - \pi_k\pi_\ell| < \infty$.

These assumptions are commonly used in the literature and are known to hold for frequently encountered sampling designs, see e.g., Robinson and Särndal (1983) and Breidt and Opsomer (2000). The non-informativeness assumption means that, conditionally on the auxiliary variables, the sample selection indicators are independent of the survey variable; see, e.g., Pfeffermann and Sverchkov (2009). Part (a) of (H1) requires that the sample sizes $\{n_v\}_{v\in\mathbb{N}}$ increase at the same rate as the population sizes $\{N_v\}_{v\in\mathbb{N}}$. Part (b) requires that both the first and second-order inclusion probabilities are bounded away from zero for the sequence of sampling designs $\{\mathcal{P}_v\}_{v\in\mathbb{N}}$. Finally, Part (c) states that the sampling covariances decrease to zero with a rate of at least $\mathcal{O}\left(n_v^{-1}\right)$.

We also assume that: i) the regression function m is continuous; ii) the distribution of the predictors $\mathbb{P}_{\mathbf{x}}$ is supported on $Supp(\mathbb{P}_{\mathbf{x}})$, a compact subset the unit cube $[0;1]^{p_v}$; iii) the residuals $\{\epsilon_k\}_{k\in U}$ have compact support. Under these assumptions, note that the survey variable Y is almost surely bounded, taking values in a set denoted $[a_y\ ;\ b_y]$, with $a_y < b_y$.

Under Assumption (H1) and the fact that the survey variable Y is bounded, it can be shown (Robinson and Särndal, 1983; Breidt and Opsomer, 2000) that there exists a positive constant C such that

$$n_v \mathbb{E}[\widehat{\mu}_{\pi,v} - \mu_v]^2 \leq C,$$

which implies that

$$\lim_{v\to\infty} \mathbb{E}[\widehat{\mu}_{\pi,v} - \mu_v]^2 = 0.$$

That is, the sequence of (complete data) Horvitz-Thompson estimators $\{\widehat{\mu}_{\pi,v}\}_{v\in\mathbb{N}}$ is *mean-square consistent* for μ_v.

4.2 Asymptotic properties of the regression tree imputed estimator

Before establishing the theoretical properties of $\widehat{\mu}_{tree}$ in a more general setting, we describe two simple yet rather unrealistic settings. For both settings, we assume that $\sum_{k\in S} d_k = N$, which holds, for instance, in the case of simple random sampling without replacement. (i) Suppose that $y_k = C$ for all $k \in U$. Then, $\widehat{\mu}_{tree}$ is a perfect estimator of μ. That is, $\widehat{\mu}_{tree} = \mu$ for all S. As a result, $\mathbb{E}[\widehat{\mu}_{tree} - \mu]^2 = 0$. (ii) Suppose that $y_k = C + \epsilon_k$, $k \in U$ for some unknown parameter C, such that $\mathbb{E}(\epsilon_k) = 0$ and $\mathbb{V}(\epsilon_k) = \sigma^2$. In addition, we assume that the tree predictor has the X-property. It follows that $\mathbb{E}\left[\widehat{\mu}_{tree} - \mu\right] = 0$ and $\mathbb{V}(\widehat{\mu}_{tree} - \mu|\boldsymbol{X}) = \sigma^2 \sum_{k\in S_r} w_k^2 > 0$, where the weights $\{w_k\}_{k\in S_r}$ satisfy $N^{-1}\sum_{k\in S_r} w_k y_k = \widehat{\mu}_{tree}$. In this setting, $\widehat{\mu}_{tree}$ remains unbiased but is no longer a perfect estimator of μ. For more general settings, the exact derivation of the bias or the variance is more challenging. To establish the asymptotic properties of the tree imputed estimator based on the CART criterion, additional notations and assumptions are needed. They are described next.

For a fixed $v \in \mathbb{N}$, let $C^1([0;1]^{p_v}, \mathbb{R})$ be the space of continuously differentiable functions defined on $[0;1]^{p_v}$ taking values in $\mathbb{R}$, and let $\mathcal{A}_v$ be the class of additive $C^1([0;1]^{p_v}, \mathbb{R})$-functions defined as

$$\mathcal{A}_v := \left\{ g_v(\mathbf{x}) = \sum_{j=1}^{p_v} g_j(x_j),\ g_j \in C^1([0;1], \mathbb{R}),\ j = 1, 2, \ldots, p_v \right\}.$$

Let $||\cdot||_{TV}$ denote the total variation norm for elements g_v in $C^1([0;1]^{p_v}, \mathbb{R})$ defined as $||g_v||_{TV} := \int_{[0;1]^{p_v}} ||\nabla g(\mathbf{x})||_1 d\mathbf{x}$, where ∇g_v denotes the gradient of g_v and $||\cdot||_1$ the 1-vector norm of $\mathbb{R}^{p_v}$ defined by $||\mathbf{x}||_1 := \sum_{j=1}^{p_v} |x_j|$. If $g_v \in \mathcal{A}_v$, then $||g_v||_{TV} = \sum_{j=1}^{p_v} \int_{[0;1]} |g'_j(x_j)| dx_j$, with ${g_v}'$ denoting the derivative of a function g_v defined on $\mathbb{R}$. In the case of a linear function g represented as $g_v(\mathbf{x}) := \mathbf{x}^\top \boldsymbol{\beta}_v$, this reduces to $||g_v||_{TV} = ||\boldsymbol{\beta}_v||_1$. Lastly, for real-valued functions defined on $\mathbb{R}^{p_v}$, we denote by $||g||_\infty := \sup_{\mathbf{x}\in\mathbb{R}^{p_v}} |g(\mathbf{x})|$ the sup-norm.

Result 4.1. *Consider a sequence of tree imputed estimators $\{\widehat{\mu}_{tree,v}\}_{v\in\mathbb{N}}$ based on the CART criterion with maximal depths $\{K_v\}_{v\in\mathbb{N}}$. Assume that (H1) holds and that the sequence of regression functions $\{m_v\}_{v\in\mathbb{N}}$ and the sequence of trees $\{\widehat{\mu}_{tree,v}\}_{v\in\mathbb{N}}$ satisfy:*

1. *$m_v \in \mathcal{A}_v$ for all $v \in \mathbb{N}$ and $\sup_{v\in\mathbb{N}} ||m_v||_\infty < \infty$.*
2. *$\lim_{v\to\infty} K_v = +\infty$, $\lim_{v\to\infty} \dfrac{||m_v||_{TV}}{\sqrt{K_v}} = 0$ and $\lim_{v\to\infty} \dfrac{2^{K_v} \log^2(n_{r,v}) \log(n_{r,v} p_v)}{n_{r,v}} = 0$.*

Then, the sequence of tree imputed estimators $\{\widehat{\mu}_{tree,v}\}_{v\in\mathbb{N}}$ *satisfies:*

$$\lim_{v\to\infty} \mathbb{E}\left[\widehat{\mu}_{tree,v} - \mu_v\right]^2 = 0.$$

Condition 1 assumes that the regression function is continuously differentiable, additive, and bounded. The first part of Condition 2 assumes that, as the sample and population sizes increase to infinity, the depth of the tree increases to infinity. The rate of divergence should be faster than the square total variation norm of the regression function. In the special case where m_v is a linear function with coefficients $\boldsymbol{\beta}_v$, this condition is automatically satisfied if the sup-norm of the regression function is uniformly bounded and $\lim_{v\to\infty} K_v = +\infty$. The second part of Condition 2 imposes a tradeoff between the growth rate of the depth, the sample size, and the number of predictors. In particular, the number of predictors can grow sub-exponentially fast, provided that the sparsity conditions imposed on the total variation and sup norm are satisfied. These additional conditions are enough to guarantee the convergence of the sequence of regression function estimators $\{m_v\}_{v\in\mathbb{N}}$ towards the regression function in L^2, as shown by Klusowski and Tian (2024); we refer the reader to Klusowski and Tian (2024) for a more thorough discussion about these conditions.

4.3 Asymptotic properties of the random forest imputed estimator

4.3.1 From finite to infinite forests

In this section, we use the concept of infinite random forest predictor, which will prove useful for (a) establishing the theoretical properties of imputed estimators based on random forests, and (b) deriving variance estimators in Section 5. For a fixed $v \in \mathbb{N}$, the infinite random forest predictor is defined by

$$\widehat{m}_{rf,v}^{(\infty)} := \mathbb{E}_\Theta\left[\widehat{m}_{rf,v}^{(B)}\right],$$

where $\mathbb{E}_\Theta$ denotes the expectation with respect to $\mathbb{P}_\Theta$. In practice, $\widehat{m}_{rf,v}^{(\infty)}$ cannot be computed. It is called an infinite forest predictor because, by the strong law of large numbers,

$$\lim_{B\to\infty} \widehat{m}_{rf,v}^{(B)} = \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \widehat{m}_{tree,v}^{(b)} \overset{a.s.}{=} \widehat{m}_{rf,v}^{(\infty)},$$

where the limit is taken in the almost sure sense. We define the infinite forest imputed estimator as

$$\widehat{\mu}_{rf,v}^{(\infty)} := \frac{1}{N}\left(\sum_{k\in S_r}\frac{y_k}{\pi_k} + \sum_{k\in S_m}\frac{\widehat{m}_{rf,v}^{(\infty)}(\mathbf{x}_k)}{\pi_k}\right). \tag{11}$$

Since an imputed forest estimator can be expressed as an average of tree imputed estimators (see relation (9) in Section 3.2), it follows from the strong law of large numbers that

$$\lim_{B\to\infty}\widehat{\mu}_{rf,v}^{(B)} \overset{a.s.}{=} \mathbb{E}_\Theta\left[\widehat{\mu}_{rf,v}^{(B)}\right] = \widehat{\mu}_{rf,v}^{(\infty)}. \tag{12}$$

Even though the infinite forest imputed estimator cannot be computed in practice, it is possible to approach it with a finite forest imputed estimator based on a large number of trees B_v. Moreover, the additional stability pertaining to $\widehat{\mu}_{rf,v}^{(\infty)}$ will be useful for deriving the theoretical properties of imputed estimators based on random forests.

Proposition 4.1. *Consider sequences of finite* $\{\widehat{\mu}_{rf,v}^{(B)}\}_{v\in\mathbb{N}}$ *and infinite* $\{\widehat{\mu}_{rf,v}^{(\infty)}\}_{v\in\mathbb{N}}$ *random forest imputed estimators. The following results hold:*

1. *For all* $b = 1,\ldots,B_v$, $\mathbb{E}[\widehat{\mu}_{rf,v}^{(B)} - \mu_v]^2 \leqslant \mathbb{E}[\widehat{\mu}_{tree,v}^{(b)} - \mu_v]^2$, *with equality if and only if either* $B = 1$ *or all trees of the forest are equal with probability one.*

2. *There exists a constant* $C > 0$, *independent of* B, *such that*

$$0 \leqslant \mathbb{E}[\widehat{\mu}_{rf,v}^{(B)} - \mu_v]^2 - \mathbb{E}[\widehat{\mu}_{rf,v}^{(\infty)} - \mu_v]^2 \leqslant \frac{C}{B_v}, \quad \textit{for all} \quad B_v \geqslant 1.$$

As a consequence, $\mathbb{E}[\widehat{\mu}_{rf,v}^{(\infty)} - \mu_v]^2 \leqslant \mathbb{E}[\widehat{\mu}_{rf,v}^{(B)} - \mu_v]^2 \leqslant \mathbb{E}[\widehat{\mu}_{tree,v}^{(b)} - \mu_v]^2$, *for all* $b = 1,...,B_v$.

Point (2) in Proposition 4.1 reveals that the mean squared error of infinite forests is, at most, equal to the mean squared error of finite forests. It follows that infinite forests are more efficient than finite forests. Proposition 4.1 also implies that the difference between the mean square errors of the infinite forests and a random forest with B_v trees decreases to 0 as B_v increases.

4.3.2 Consistency of the random forest imputed estimators

We begin by considering the uniform random forests described in Section 3.2. An important part of the proof is based on the idea that the forests under consideration are, in some sense,

large and stable. That is, we assume that, without any requirement on the rate, that the number of trees is strictly increasing, which implies that for $v_1 < v_2$ positive integers, the number of trees B_{v_1} in $\widehat{m}_{rf}^{(B_{v_1})}$ used to impute in S_{v_1} is strictly smaller than the number of trees B_{v_2} used to obtain $\widehat{m}_{rf}^{(B_{v_2})}$ in S_{v_2}. In other words, $v_1 < v_2$ implies that $B_{v_1} < B_{v_2}$. As a result, $\lim_{v\to\infty} B_v = +\infty$.

Result 4.2. *Consider a sequence of uniform forest imputed estimators $\{\widehat{\mu}_{urf,v}^{(B)}\}_{v\in\mathbb{N}}$ based on trees with depths $\{K_v\}_{v\in\mathbb{N}}$. Suppose that assumption (H1) holds. We also assume that:*

(1) The sequence of regression functions $\{m_v\}_{v\in\mathbb{N}}$ satisfies $\sup_{v\in\mathbb{N}}||m_v||_\infty < \infty$.

(2) The depths $\{K_v\}_{v\in\mathbb{N}}$ increase as v increases such that the following conditions are satisfied:

$$(a)\ \lim_{v\to\infty} p_v\left(1-\frac{1}{4p_v}\right)^{K_v} = 0, \qquad \text{and} \qquad (b)\ \lim_{v\to\infty}\frac{2^{K_v}}{n_v} = 0.$$

(3) The number of trees in the forest increases, i.e., $\lim_{v\to\infty} B_v = +\infty$.

Then, the uniform random forest imputed estimator $\{\widehat{\mu}_{urf,v}^{(B)}\}_{v\in\mathbb{N}}$ satisfies:

$$\lim_{v\to\infty}\mathbb{E}[\widehat{\mu}_{urf,v}^{(B)} - \mu_v]^2 = 0.$$

The condition given in Part (1) of Result 4.2 follows from sufficient conditions for the mean square consistency in high-dimensional settings of $\{\widehat{m}_{urf,v}^{(B)} - \widehat{m}_v\}_{v\in\mathbb{N}}$ towards 0; see the Appendix. If the number of predictors is fixed, this condition reduces to the conditions found in (Scornet, 2016, Corrolary 3.1). Condition (a) of Part (2) ensures the diameters of each node decrease to 0 as v increases. It is satisfied, for instance, if K_v increases fast enough compared to p_v. Condition (b) is sufficient to ensure that the probability of having an empty leaf converges to 0. Result 4.3 below extends the consistency of regression trees imputed estimators to Breiman's random forests.

Result 4.3. *Consider a sequence of random forest imputed estimators $\{\widehat{\mu}_{brf,v}^{(B)}\}_{v\in\mathbb{N}}$ based on Breiman's algorithm. The trees from the forests have maximal depths $\{K_v\}_{v\in\mathbb{N}}$. Suppose that assumption (H1) holds. We also assume that:*

1. *The sequence of regression functions $\{m_v\}_{v\in\mathbb{N}}$ satisfies $m_v \in \mathcal{A}_v$ and $\sup_{v\in\mathbb{N}}||m_v||_\infty < \infty$.*

2. *The sequence of random forests* $\{\widehat{m}^{(B)}_{brf,v}\}_{v\in\mathbb{N}}$ *satisfies*

$$(a)\ \lim_{v\to\infty} K_v = +\infty, \qquad (b)\ \lim_{v\to\infty} \frac{\sqrt{p_v}||m_v||_{TV}}{\sqrt{p_{0v}}\sqrt{K_v}} = 0, \qquad (c)\ \lim_{v\to\infty} \frac{2^{K_v}\log^2(a_v)\log(a_v\ p_v)}{a_v} = 0.$$

3. *The number of trees* B_v *in the forest increases i.e.,* $\lim_{v\to\infty} B_v = +\infty$.

Then, the random forest imputed estimator $\{\widehat{\mu}^{(B)}_{brf,v}\}_{v\in\mathbb{N}}$ *based on Breiman's algorithm satisfies*

$$\lim_{v\to\infty} \mathbb{E}[\widehat{\mu}^{(B)}_{brf,v} - \mu_v]^2 = 0.$$

The conditions obtained on the parameters of $\{\widehat{m}^{(B)}_{brf,v}\}_{v\in\mathbb{N}}$ are similar to those presented in Result 4.1. The impact of the number of predictors, p_{0v}, to be selected randomly at each split, also appears in Condition 2. As noted by Klusowski and Tian (2024), this condition allows for choosing p_{0v} negligible with respect to p_v, provided that the depth K_v increases fast enough, and that the regression function is sparse enough. However, these results are asymptotic, and special care is needed to choose this parameter in practice in the context of imputation. This aspect is further discussed in Section 7.

Remark 4.1. *In Result 4.2 and Result 4.3, we require that the number of trees* B_v *increases to infinity, i.e.,* $\lim_{v\to\infty} B_v = +\infty$*, without rate requirement. This condition is indeed sufficient to ensure the following implication:*

$$\lim_{v\to\infty} \mathbb{E}[\widehat{\mu}^{(\infty)}_{rf,v} - \mu_v]^2 = 0 \quad \Longrightarrow \quad \lim_{v\to\infty} \mathbb{E}[\widehat{\mu}^{(B)}_{rf,v} - \mu_v]^2 = 0,$$

where $\{\widehat{\mu}_{rf,v}\}_{v\in\mathbb{N}}$ *denotes any sequence of random forest estimators. Clearly, the reverse implication always holds. However, to ensure that the convergence rates of both finite and infinite imputed estimators are the same, the stronger requirement of* $\lim_{v\to\infty} n_v/B_v = 0$ *would be needed.*

5 Variance estimation

In this section, we study the problem of variance estimation in the context of imputed data through regression trees and random forests. We start by describing the naive variance estimator, which is obtained by applying a complete data variance estimation procedure to

the pseudo-values, $\widetilde{y}_k = r_k y_k + (1 - r_k)\widehat{m}_{rf}^{(B)}(\mathbf{x}_k)$, obtained after imputation. This leads to

$$\widehat{V}_{naive} := \frac{1}{N_v^2} \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widetilde{y}_k}{\pi_k} \frac{\widetilde{y}_\ell}{\pi_\ell}, \tag{13}$$

where $\Delta_{k\ell} := \pi_{k\ell} - \pi_k \pi_\ell$. In general, the use of (13) may lead to a severe underestimation of the total variance of $\widehat{\mu}_{rf}^{(B)}$. This is illustrated empirically in Section 6.2. To derive variance estimators that account for sampling and nonresponse, we first decompose the total variance of $\widehat{\mu}_{rf}^{(B)}$.

Proposition 5.1. *Consider sequences of finite* $\{\widehat{\mu}_{rf,v}^{(B)}\}_{v \in \mathbb{N}}$ *and infinite* $\{\widehat{\mu}_{rf,v}^{(\infty)}\}_{v \in \mathbb{N}}$ *forest estimators. We have*

$$\mathbb{V}\left(\widehat{\mu}_{rf,v}^{(B)} - \mu_v\right) = \mathbb{V}\left(\widehat{\mu}_{rf,v}^{(\infty)} - \mu_v\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf,v}^{(B)}\right)\right], \tag{14}$$

where $\mathbb{V}_\Theta$ *denote the variance operator with respect to* $\mathbb{P}_\Theta$*. Furthermore, there exists* $C > 0$ *such that*

$$\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf,v}^{(B)}\right)\right] \leqslant \frac{C}{B_v}. \tag{15}$$

It follows from Proposition 4.1 and (14), that the contribution of $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf,v}^{(\infty)}\right)\right]$ to the total variance $\mathbb{V}(\widehat{\mu}_{rf,v}^{(B)})$ is negligible provided that $n_v/B_v = o(1)$. Proposition 5.1 suggests that the contribution of the randomization variance can be made arbitrarily small by choosing a large value of B_v. Empirical results in Section 7 suggest that the contribution $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf,v}^{(\infty)}\right)\right]$ is small for moderate values of B_v. As a result, in Sections 5.1 and 5.2, we omit the term $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf,v}^{(\infty)}\right)\right]$ from the computations.

5.1 Variance estimation based on a first-order Taylor expansion

An estimator of the variance of $\widehat{\mu}_{rf}^{(B)}$ can be obtained through the so-called reverse approach of Fay (1991) and Shao and Steel (1999); see also Kim and Rao (2009) and Haziza and Vallée (2020). This approach leads to the following decomposition of the total variance of $\widehat{\mu}_{rf}^{(B)}$:

$$\begin{aligned} \mathbb{V}\left(\widehat{\mu}_{rf}^{(B)} - \mu | \boldsymbol{r}\right) &= \mathbb{E}\left[\mathbb{V}\left(\widehat{\mu}_{rf}^{(B)} | \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{X}\right) | \boldsymbol{r}, \boldsymbol{X}\right] + \mathbb{V}\left[\mathbb{E}\left(\widehat{\mu}_{rf}^{(B)} - \mu | \boldsymbol{r}, \boldsymbol{y}, \boldsymbol{X}\right) | \boldsymbol{r}, \boldsymbol{X}\right] \\ &:= V_1 + V_2, \end{aligned} \tag{16}$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$. As noted by various authors (Shao and Steel, 1999; Haziza and Vallée, 2020), the contribution of the second term on the right-hand-side of (16) to the total variance is negligible if the sampling fraction n/N is negligible. In the sequel, we assume that n/N is negligible, which is commonly encountered in practice.

Using a first-order Taylor expansion, an estimator of V_1 in (16) is given by

$$\widehat{V}_1 := \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\xi}_k^{(rf,B)}}{\pi_k} \frac{\widehat{\xi}_\ell^{(rf,B)}}{\pi_\ell}, \tag{17}$$

where, for $k \in S$, we have

$$\begin{aligned} \widehat{\xi}_k^{(rf,B)} &:= \frac{1}{B} \sum_{b=1}^{B} \widehat{\xi}_k^{(tree,b)} \\ &:= r_k y_k + (1 - r_k)\, \widehat{m}_{rf}^{(B)}(\mathbf{x}_k) + \frac{1}{B} \sum_{b=1}^{B} \psi_k^{(b)} \frac{\widehat{N}_b\left(\mathbf{x}_k, S_m\right)}{\widehat{N}_b\left(\mathbf{x}_k, S_r\left(\Theta_b\right)\right)} \cdot \left(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)\right), \end{aligned} \tag{18}$$

where the notations $\widehat{N}_b\left(\mathbf{x}_k, S_m\right)$ and $\widehat{N}_b\left(\mathbf{x}_k, S_r\left(\Theta_b\right)\right)$ are defined in (5), and $S_r\left(\Theta_b\right)$ denotes the set of elements selected in the b-th subsample.

We point out that the derivation of $\widehat{V}_1$ was made conditionally on the partition of the predictor space. This is a simplification of reality as the partitions vary from one sample to another. Also, for small values of n_0, the estimator $\widehat{V}_1$ may suffer from serious underestimation, largely due to overfitting. Indeed, the residuals, $y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)$, in (18), can be made artificially small for small values of n_0. An extreme case arises when all sample residuals are equal to zero, in which case $\widehat{V}1 = \widehat{V}naive$, leading to severe underestimation. A similar issue was encountered by Opsomer and Miller (2005) for selecting the bandwidth of model-assisted estimators based on local polynomials, and in Dagdoug et al. (2023b) in the context of model-assisted estimation based on random forests. To overcome this problem, we propose a novel variance estimator based on a K-fold cross-validation procedure in Section 5.2, which extends the K-fold variance estimation procedure of Dagdoug et al. (2023b). The reader is referred to Arlot and Celisse (2010) for a comprehensive overview of cross-validation procedures.

Remark 5.1. *For a single deterministic regression tree $\widehat{m}_{tree}$, the linearized variable given*

by (18) reduces to

$$\widehat{\xi}_k^{(tree)} = r_k y_k + (1 - r_k)\, \widehat{m}_{tree}(\mathbf{x}_k) + \frac{\widehat{N}\left(\mathbf{x}_k, S_m\right)}{\widehat{N}\left(\mathbf{x}_k, S_r\left(\Theta_b\right)\right)} \cdot \left(y_k - \widehat{m}_{tree}(\mathbf{x}_k)\right), \qquad k \in S.$$

5.2 A variance estimator based on a K-fold cross-validation procedure

To circumvent the overfitting issue, we propose a new variance estimator based on a K-fold cross-validation procedure. The proposed variance estimator is identical to $\widehat{V}_1$ in (17), except that the residuals $\widehat{\epsilon}_k \equiv \left(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k)\right), k \in S_r$, in (18) are replaced with residuals constructed through a K-fold cross-validation procedure. Most often the value of K is set to $K = 3; 5; 10$. More specifically, we proceed as follows: We start by replicating the set of respondents S_r, K times. Each of these K datasets is then split into two disjoint subsets $S_{r,train}^{(j)}$ and $S_{r,test}^{(j)}$ of respective sizes $n_{train} := n_r \times (K-1)/K$, which we assume to be an integer for simplicity, and $n_{test} := n - n_{train}$, for $j = 1, 2, \ldots, K$. Then, for each of the K partitions, a random forest estimator of m is fitted on $S_{r,train}^{(j)}$. The B trees fitted on $S_{r,train}^{(j)}$ are then used to make predictions on $S_{r,test}^{(j)}$, individually; that is, we do not average these predictions to obtain the predictions of the forest, but we store the predictions from each of these individual trees. Since each $\{S_{r,train}^{(j)}, S_{r,test}^{(j)}\}_{j=1,\ldots,K}$ leads to a partition of S_r, each of the B trees residuals computed on $S_{r,test}^{(j)}$ are uniquely defined. We thus obtain a set $\{\widehat{\epsilon}_k^{(cv,b)} \ ; \ k \in S_r, \ b \in \{1, \ldots, B\}\}$ of $B \times n_r$ tree residuals. This leads to the proposed variance estimator of V_1:

$$\widehat{V}_1^{(cv)} := \sum_{k \in S} \sum_{\ell \in S} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{\widehat{\xi}_k^{(cv)}}{\pi_k} \frac{\widehat{\xi}_\ell^{(cv)}}{\pi_\ell}, \tag{19}$$

where

$$\widehat{\xi}_k^{(cv)} := r_k y_k + (1 - r_k)\, \widehat{m}_{rf}^{(B)}(\mathbf{x}_k) + \frac{1}{B} \sum_{b=1}^{B} \psi_k^{(b)} \frac{\widehat{N}_b\left(\mathbf{x}_k, S_m\right)}{\widehat{N}_b\left(\mathbf{x}_k, S_r\right)} \cdot \widehat{\epsilon}_k^{(b,cv)}, \qquad k \in S. \tag{20}$$

6 Simulation studies

In this section, we conduct a simulation study to compare the performance of several imputation procedures in terms of bias and efficiency and to compare the performance of several variance estimation procedures in terms of bias and coverage rate of normal-based confidence intervals.

6.1 Point estimation

We repeated $R = 5,000$ iterations of the following process:

(i) A finite population of size $N = 5,000$ was generated. The population consisted of a set of $p = 90$ predictors $X_1, \ldots, X_{90}$, and 5 survey variables $Y_1, \ldots, Y_5$. To generate the X-variables, we considered two scenarios: (i) The predictors were generated independently from a normal distribution with mean equal to 5 and variance equal to 1. (ii) The predictors were generated from a multivariate normal distribution with a mean vector equal to $5 \times \mathbf{1}^\top$ and variance-covariance matrix whose diagonal elements were equal to 1 and the off-diagonal elements were equal to 0.7, where $\mathbf{1}$ denotes the vector of ones. That is, the predictors were correlated. Given the values of $X_1, \ldots, X_{90}$, we generated 5 survey variables according to the following models:

$$
\begin{aligned}
Y_1 &= 2 + X_1 + 3X_2 + 4X_5 + \mathcal{N}(0, 5), \\
Y_2 &= 10^{-3} X_1^6 X_2^3 + \mathcal{N}(0, 1), \\
Y_3 &= 1.5 + \cos(X_1 + X_2 + X_3 + X_4) + \mathcal{N}\left(0, 10^{-2}\right), \\
Y_4 &= 2 + \mathbb{1}_{\{X_1 > 7\}} - \mathbb{1}_{\{X_1 < 4\}} + 2\mathbb{1}_{\{X_4 > 6\}} + \mathcal{N}(0, 1), \\
Y_5 &= 2 + X_1 + 10 \exp\left(2\mathbb{1}_{\{X_5 > 5\}} - \mathbb{1}_{\{X_5 < 6\}}\right).
\end{aligned}
$$

Note that the survey variables $Y_1, \ldots, Y_5$ were generated using a subset of the first five predictors $X_1, \ldots, X_5$. We were interested in estimating the population means of $Y_1, \ldots, Y_5$, denoted by $\mu_1, \ldots, \mu_5$, respectively.

(ii) From the finite population generated in Step (i), a sample of size $n = 250$ was selected according to simple random sampling without replacement.

(iii) In each sample, the response indicators $r_k,\ k \in S$, were independently generated according to a Bernoulli distribution with probability

$$
p_k = \text{logit}\left(0.15 \times \{-30 + X_1 + X_2 + X_3 + X_4 + 2X_5\}\right). \tag{21}
$$

This led to a response rate approximately equal to 50%.

(iv) The missing values in each sample were imputed by five imputation procedures:

(1) Deterministic linear regression imputation;

(2) Regression tree (CART) imputation, with $n_0 = 10$ and a complexity parameter[1] `cp` $= 0.01$. The package `Rpart` (Therneau and Atkinson, 2022) was used for the implementation of CART imputation.

(3) Random forest (RF) imputation with $B = 1\ 000$ trees, $n_0 = 10$ elements in each terminal node and $p_0 = p$. We used the bootstrap as the resampling algorithm. We implemented RF imputation using the package `Ranger` (Wright and Ziegler, 2017).

(4) Nearest neighbor (NN) imputation.

(5) K-nearest neighbour (KNN) imputation with $K = 5$. The package `caret` (Kuhn, 2022) was used for the implementation of both NN and KNN.

Because we were interested in understanding the impact of the number of predictors on the behavior of the resulting imputed estimator, we considered two scenarios: (a) the case where only the first 5 predictors were included in each model; b) the case where the 90 predictors were included.

(v) In each completed data set and for each imputation procedure, we computed the estimators $\widehat{\mu}_{\widehat{m}}$ given by (3).

As a measure of bias, we used the Monte-Carlo percent relative bias (RB), defined as

$$RB(\widehat{\mu}_{\widehat{m},j}) := 100 \times \frac{1}{R}\sum_{r=1}^{R}\frac{\left(\widehat{\mu}_{\widehat{m},j}^{(r)} - \mu_j^{(r)}\right)}{\mu_j^{(r)}}, \qquad j = 1, 2, \ldots, 5,$$

where $\hat{\mu}_{\widehat{m},j}^{(r)}$ denotes an estimator of μ_j at the r-th iteration, $r = 1, \ldots, R$. As a measure of relative efficiency (RE) with respect to the Horvitz-Thompson estimator, we used

$$RE(\hat{\mu}_{\widehat{m},j}) := 100 \times \frac{\sum_{r=1}^{R}\left(\hat{\mu}_{\widehat{m},j}^{(r)} - \mu_j^{(r)}\right)^2}{\sum_{r=1}^{R}\left(\hat{\mu}_{j\pi}^{(r)} - \mu_j^{(r)}\right)^2}, \qquad j = 1, 2, \ldots, 5.$$

The results for $p = 5$ with independent and correlated predictors are given in Table 1 and Table 2, respectively. In both cases, LR was, as expected, the most efficient estimator for the

[1]The complexity parameter `cp` is a parameter available in the `Rpart` package, whereby, as per the `Rpart` documentation, "any split that does not decrease the overall lack of fit by a factor of `cp` is not attempted".

Survey variable	MC measure	Imputed estimators				
		LR	NN	5NN	CART	RF
Y_1	RB	0.0	0.5	0.7	0.6	0.4
	RE	158	222	227	220	186
Y_2	RB	-1.4	-2.8	-3.1	3.4	1.0
	RE	137	123	121	168	124
Y_3	RB	0.4	1.1	1.3	-0.4	-0.1
	RE	213	205	206	217	178
Y_4	RB	-0.1	-0.3	-0.3	0.2	0.1
	RE	197	229	202	187	181
Y_5	RB	-2.1	-2.1	-2.6	1.0	0.0
	RE	164	157	153	105	104

Table 1: Monte Carlo Simulation Results for $p = 5$ and independent predictors.

Survey variable	MC measure	Imputed estimators				
		LR	NN	KNN	CART	RF
Y_1	RB	0.0	0.4	0.8	0.9	0.5
	RE	137	178	185	209	160
Y_2	RB	-19.2	-1.0	-1.1	7.4	1.3
	RE	368	105	103	179	108
Y_3	RB	-0.1	-0.4	-0.8	-0.2	-0.5
	RE	247	188	200	239	196
Y_4	RB	-1.7	0.0	0.4	0.3	0.1
	RE	213	234	202	193	186
Y_5	RB	-10.0	-0.9	-1.0	2.8	0.1
	RE	310	124	123	113	103

Table 2: Monte Carlo Simulation Results for $p = 5$ and correlated predictors.

variable Y_1 with a value of RE of about 158% for $p = 5$ independent predictors. RF performed better than the other procedures with a value of RE equal to 186% for $p = 5$ independent predictors and 160% for $p = 5$ correlated predictors. For the survey variables $Y_2, \ldots, Y_5$, LR was generally biased, as expected. The biases were larger in the case of correlated predictors. RF, on the other hand, exhibited negligible bias across all scenarios. In terms of RE, RF was either comparable to the best procedure or the most efficient overall. The procedures NN,

5NN, and CART were also efficient in most scenarios. However, in some scenarios, NN and 5NN exhibited a slight bias. This is likely due to the curse of dimensionality, e.g., Abadie and Imbens (2006) and Yang and Kim (2019). RF outperformed CART in all the scenarios.

Survey variable	MC measure	Imputed estimators				
		LR	NN	5NN	CART	RF
Y_1	RB	0.0	1.5	1.6	0.8	0.9
	RE	305	588	536	259	263
Y_2	RB	-1.5	2.4	4.0	2.7	3.2
	RE	286	286	256	158	150
Y_3	RB	0.5	-0.9	-1.1	-1.3	-1.7
	RE	485	396	315	259	231
Y_4	RB	-0.1	1.2	1.3	0.4	0.6
	RE	461	386	293	193	177
Y_5	RB	-1.9	8.6	9.4	1.0	0.2
	RE	348	52	8457	104	104

Table 3: Monte Carlo Simulation Results for $p = 90$ and independent predictors.

Survey variable	MC measure	Imputed estimators				
		LR	NN	5NN	CART	RF
Y_1	RB	0.0	0.8	1	1.1	0.9
	RE	249	231	226	237	193
Y_2	RB	-23.7	0.8	0.8	8.5	2.4
	RE	790	112	107	210	111
Y_3	RB	0.0	0.1	-0.4	0.1	-0.1
	RE	615	275	226	268	205
Y_4	RB	-1.6	0.8	1.1	0.6	0.5
	RE	519	293	233	213	190
Y_5	RB	-8.8	2.3	2.4	2.5	0.1
	RE	507	181	155	111	102

Table 4: Monte Carlo Simulation Results for $p = 90$ and correlated predictors.

The results for $p = 90$ with independent and correlated predictors are given in Table 3 and Table 4, respectively. In most scenarios, RF exhibited negligible bias and was the most efficient. For the survey variable Y_1, RF outperformed LR with independent and correlated

predictors. This is not surprising as the performance of linear regression imputation tends to deteriorate as the dimension of the **x**-vector increases. Again, NN and 5NN suffered from the curse of dimensionality. This was especially evident in the case of independent predictors, where NN and KNN displayed a relative bias equal to 8.6% and 9.4%, respectively. Comparing the results in Table 1 and Table 3, it is worth mentioning that, unlike the other estimators, the performance of RF was only moderately impacted by the dimension of the **x**-vector, which suggests that RF remain efficient even in a high-dimensional setting.

6.2 Performance of variance estimators

In this section, we investigate the performance of both the linearized variance estimator proposed in Section 5.1 and the novel variance estimator based on a K-fold cross-validation procedure proposed in Section 5.2. We used the same setup described in Section 6.1. As a measure of bias of a variance estimator $\widehat{V}$, we computed its Monte-Carlo percent relative bias (RB) given by

$$RB(\widehat{V}) := 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{\widehat{V}^{(r)} - V_{MC}\left(\widehat{\mu}\right)}{V_{MC}\left(\widehat{\mu}\right)},$$

with $V_{MC}\left(\widehat{\mu}\right)$ denoting the Monte-Carlo variance of $\widehat{\mu}$. We also computed the Monte Carlo coverage rate of 95% normal-based confidence intervals of the form

$$IC_r\left(\widehat{\mu}^{(r)}, \widehat{V}^{(r)}\right) := \left[\widehat{\mu}^{(r)} - 1.96 \times \sqrt{\widehat{V}^{(r)}}\ ;\ \widehat{\mu}^{(r)} + 1.96 \times \sqrt{\widehat{V}^{(r)}}\right].$$

The Monte-Carlo coverage rate is then defined as

$$\text{Coverage}\left(\widehat{\mu}^{(r)}, \widehat{V}^{(r)}\right) := \frac{100}{R} \sum_{r=1}^{R} \mathbb{1}_{\mu \in \left\{IC_r\left(\widehat{\mu}_t^{(r)}, \widehat{V}^{(r)}\right)\right\}}.$$

As in Section 6.1, the sample size n was set to 250, which corresponds to a sampling fraction, n/N, equal to 5%. This can be viewed as a small sampling fraction. As in Section 6.1, the sample size n was set to 250, corresponding to a sampling fraction $n/N = 5\%$. This is considered a small sampling fraction.

6.2.1 Variance estimation: Imputation through regression tree

Results for the case of $p = 5$ independent predictors and correlated predictors are presented in Table 5 and Table 6, respectively. Results for the case of $p = 90$ with independent

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-56.4	77.9	-59.9	74.0	-66.4	65.0
	Linearized	-27.4	88.6	-18.9	89.4	-16.7	86.0
	CV	5.0	94.0	0.1	92.4	-5.8	88.2
Y_2	Naïve	-34.4	88.2	-42.5	86.0	-48.9	82.5
	Linearized	-9.8	92.0	-6.6	93.5	-2.8	94.1
	CV	6.2	93.8	4.5	94.8	4.3	95.0
Y_3	Naïve	-66.0	74.1	-68.8	72.0	-71.6	69.4
	Linearized	-33.0	88.8	-21.5	91.5	-13.1	92.6
	CV	3.6	95.1	0.3	95.0	-1.6	94.2
Y_4	Naïve	-58.9	79.0	-59.9	78.6	-63.4	76.7
	Linearized	-29.5	90.1	-18.7	92.3	-15.6	92.7
	CV	1.7	95.3	-2.4	94.8	-5.2	94.0
Y_5	Naïve	-3.9	94.0	-2.9	94.4	-7.7	93.0
	Linearized	-3.7	94.0	-2.6	94.4	-1.9	94.8
	CV	-0.7	94.4	0.4	94.7	6.6	95.5

Table 5: Monte-Carlo simulation results for tree variance estimators for $p = 5$ and independent predictors.

predictors and correlated predictors are presented in Table 7 and Table 8, respectively.

As expected, the naive variance estimator suffered from large negative biases, leading to substantial undercoverage in most scenarios. The linearized variance estimator $\widehat{V}_1$ given by (17) exhibited noticeable negative bias as well, although not as prominently as the naive variance estimator. The bias was especially appreciable for small values of n_0. For instance, for Y_1, the relative bias of the linearized variance estimator ranged between -33.5% and -3.3% for $p = 5$ correlated predictors, whereas the values of the coverage rate ranged between 88.4% and 94.1%. A similar pattern was observed in the other scenarios. As mentioned in Section 5.2, the poor behavior of (17) is most likely due to overfitting: small values of n_0 tend to produce artificially small sample residuals, which, in turn, produce variance estimates that are too small. In contrast, the proposed variance estimator based on 10-fold cross-validation procedure, performed well. For instance, for $p = 5$ correlated predictors (See Table 6) and

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naive	-46.3	82.3	-50.0	76.6	-57.0	53.2
	Linearized	-23.0	89.2	-17.5	86.4	-14.5	72.7
	CV	4.5	93.9	-0.2	89.7	-2.1	76.4
Y_2	Naive	-14.3	91.5	-17.6	91.8	-26.3	90.5
	Linearized	-3.6	92.8	0.2	94.2	-0.5	94.7
	CV	3.3	93.6	5.8	94.8	4.0	95.2
Y_3	Naive	-68.4	72.7	-71.5	70.2	-75.3	65.3
	Linearized	-33.5	88.4	-23.6	91.0	-13.8	91.4
	CV	4.7	94.8	1.6	94.8	-2.8	93.3
Y_4	Naive	-60.2	78.5	-61.1	78.0	-62.4	76.5
	Linearized	-30.0	89.9	-21.0	91.8	-14.5	92.5
	CV	1.9	94.8	-3.1	94.4	-5.5	93.8
Y_5	Naive	-3.5	94.0	-3.4	94.1	-1.3	94.5
	Linearized	-3.3	94.0	-3.2	94.1	-0.5	94.8
	CV	-1.6	94.1	-1.7	94.4	2.1	95.4

Table 6: Monte-Carlo simulation results for tree variance estimators for $p = 5$ and correlated predictors.

$n_0 = 5$, the values of relative bias ranged from -0.7% to 6.2% in the case of $Y_1, \ldots, Y_5$, and the coverage rate ranged from 93.8% to 95.3%. Similar results were obtained for $p = 90$ predictors.

6.2.2 Variance estimation: Imputation through random forests

In this section, we present the results for imputation through random forests. The forests were based on $B = 100$ trees, and the value of p_0 was set to p. The choice of p_0 is further discussed in Section 7.3. Results for $p = 5$ correlated predictors are presented in Table 9. Results for $p = 90$ predictors were very similar and are thus omitted. From Table 9, we note that the proposed variance estimator based on a 10-fold cross-validation procedure performed well, especially for $n_0 = 5$ and $n_0 = 10$. In contrast, the naive variance estimator and the linearized variance estimator suffered from substantial bias in most scenarios.

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-60.9	71.6	-61.8	70.6	-64.7	64.8
	Linearized	-44.4	80.0	-28.1	85.1	-14.5	85.6
	CV	2.5	91.7	-0.4	90.8	-1.8	88.5
Y_2	Naïve	-37.5	87.2	-41.0	86.0	-46.2	83.4
	Linearized	-17.6	91.4	-7.3	93.4	-0.5	94.5
	CV	2.4	94.7	4.4	94.9	6.9	95.3
Y_3	Naïve	-66.8	72.2	-68.7	70.7	-72.1	66.8
	Linearized	-48.7	82.0	-33.5	87.3	-20.9	89.8
	CV	6.1	94.6	1.9	93.8	-3.9	92.7
Y_4	Naïve	-58.7	78.9	-59.7	78.7	-62.0	77.4
	Linearized	-42.0	86.3	-27.4	90.8	-16.6	92.7
	CV	3.5	95.3	-1.1	94.7	-0.5	94.9
Y_5	Naïve	-4.1	94.1	-4.8	94.1	-6.8	93.6
	Linearized	-4.0	94.1	-4.5	94.2	-4.7	94.3
	CV	-1.4	94.4	-2.2	94.5	0.3	95.1

Table 7: Monte-Carlo simulation results for tree variance estimators for $p = 90$ and independent predictors.

7 Choice of hyper-parameters

Random forest algorithms require the specification of several hyper-parameters. In this section, we discuss the choice of three hyper-parameters: the number of trees B, the number of observations n_0 in each terminal node, and p_0, the number of predictors randomly selected at each split.

7.1 Choice of B

Selecting the number of trees B to be used is likely the simplest parameter to decide on: the more, the better. Indeed, choosing a large value of B leads to more efficient point estimators of a population mean; see Proposition 4.1 and Proposition 5.1. Also, it simplifies the variance estimation process, as the second term on the right-hand side of (14) can be safely omitted from the computations.

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-50.1	78.3	-50.5	74.3	-57.7	53.9
	Linearized	-36.3	83.6	-24.6	84.3	-20.3	71.8
	CV	3.2	91.8	0.8	89.4	-4.1	76.7
Y_2	Naïve	-19.8	90.6	-23.2	91	-33.1	91
	Linearized	-3.8	91.5	0.3	93.1	-1.7	94.9
	CV	6.1	92.5	6.5	93.7	2.3	95.3
Y_3	Naïve	-70.7	71.7	-72.6	69.2	-75.9	64.8
	Linearized	-50.8	82.9	-37.7	87	-24.3	89.6
	CV	4.2	94.6	-1	94.2	-7.1	92.9
Y_4	Naïve	-61.5	77.6	-61.6	77.3	-63.1	76.2
	Linearized	-44.2	85.9	-31	89.2	-22.3	90.9
	CV	1.9	94.8	-3.1	94.1	-8.8	93.1
Y_5	Naïve	-2.5	94.2	-1.7	94.4	-4	94.1
	Linearized	-2.4	94.2	-1.4	94.4	-3.6	94.2
	CV	-0.6	94.4	0.1	94.7	-1.9	94.7

Table 8: Monte-Carlo simulation results for tree variance estimators for $p = 90$ and correlated predictors.

The contribution of the randomization variance $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf}^{(B)}\right)\right]$ to the total variance $\mathbb{V}\left(\widehat{\mu}_{rf}^{(B)}\right)$ was assessed through a simulation study. The Monte-Carlo contribution of $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf}^{(B)}\right)\right]$ is given by

$$\text{Contribution}_{MC}\left(\widehat{\mu}_{rf}^{(B)}\right) := 100 \times \frac{\frac{1}{R}\sum_{r=1}^{R} V_{MC,\Theta}^{(r)}\left(\widehat{\mu}_{rf}^{(B)}\right)}{V_{MC}\left(\widehat{\mu}_{rf}^{(B)}\right)},$$

where $V_{MC}\left(\widehat{\mu}_{rf}^{(B)}\right)$ denotes the usual Monte-Carlo variance of $\widehat{\mu}_{rf}^{(B)}$ and $V_{MC,\Theta}^{(r)}\left(\widehat{\mu}_{rf}^{(B)}\right)$ denotes the Monte-Carlo conditional variance of $\widehat{\mu}_{rf}^{(B)}$ computed by, conditionally on the rth population, the r-th sample and the r-th set of respondents, resampling from $\mathbb{P}_\Theta$ a number R_Θ of iterations to compute the Monte-Carlo variance of the estimator $\widehat{\mu}_{rf}^{(B)}(\Theta)$. Results for the survey variables Y_1-Y_5 (see Section 6) are shown in Figure 1. From Figure 1, the contribution of the randomization variance decreased rapidly as B increased. With $B = 50$

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-45.2	83.1	-45.1	81.9	-48.4	77.2
	Linearized	-40.8	84.6	-33.2	85.9	-26.0	85.4
	CV	6.4	93.9	5.3	93.3	0.3	90.8
Y_2	Naïve	-10.8	91.6	-9.4	91.9	-13.8	92.0
	Linearized	-9.2	92.0	-5.4	92.6	-7.3	93.4
	CV	-0.7	93.2	2.7	93.6	0.4	94.3
Y_3	Naïve	-67.2	73.8	-67.5	73.4	-70.3	71.5
	Linearized	-60.4	78.4	-48.2	83.9	-37.6	87.6
	CV	4.2	94.6	6.8	95.3	6.1	95.2
Y_4	Naïve	-60.7	78.0	-60.7	78.4	-60.7	78.3
	Linearized	-54.5	81.3	-45.2	85.5	-34.8	88.9
	CV	6.6	95.4	4.4	95.1	1.8	95.0
Y_5	Naïve	-3.0	94.1	-3.9	94.2	-1.7	94.2
	Linearized	-2.9	94.1	-3.8	94.2	-1.5	94.2
	CV	-0.7	94.4	-1.7	94.4	0.5	94.5

Table 9: Monte-Carlo simulation results for random forest variance estimators for $p = 5$ and correlated predictors.

trees, the contribution of the randomization variance fell below 3% for all survey variables. The results of this experiment suggest that we can safely omit the randomization variance from the computations for large B, say $B = 1,000$.

Next, we provide a concentration inequality that highlights that, with high probability, the random forest imputed estimator based on a finite number of trees B can be made arbitrarily close to the (infeasible) infinite forest imputed estimator.

Proposition 7.1. *Fix $B_v \in \mathbb{N}$ and $\epsilon > 0$. The probability that the finite forest imputed estimator is not in an ϵ-neighbourhood of the infinite forest estimator is bounded by*

$$\mathbb{P}_\Theta\left(|\widehat{\mu}_{rf,v}^{(B)} - \widehat{\mu}_{rf,v}^{(\infty)}| \geqslant \epsilon\right) \leqslant 2\exp\left(\frac{-B_v\epsilon^2}{2\frac{n_{m,v}^2}{N_v^2}\left(\frac{b_y - a_y}{\min_{k\in U_v}\pi_k}\right)^2}\right), \tag{22}$$

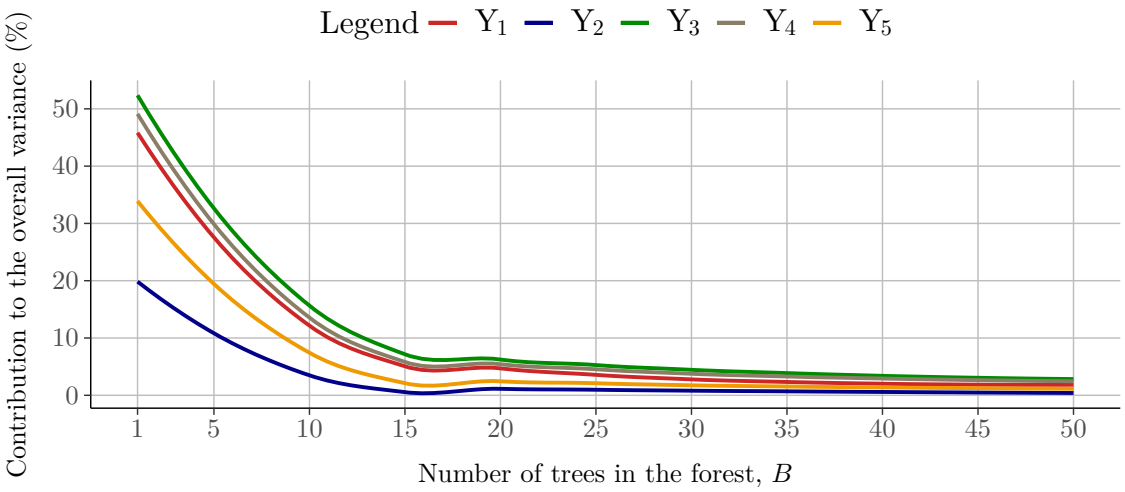

Figure 1: Contribution of the randomization variance $\mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}_{rf,j}^{(B)}\right)\right]$ to the overall variance $\mathbb{V}\left(\widehat{\mu}_{rf,j}^{(B)}\right)$ as a function of B, with $p = 5$ correlated predictors.

where $[a_y\ ;\ b_y]$ *denotes the support of* Y.

Since the bound given decreases to 0 as B increases, it follows from (22) that $\widehat{\mu}_{rf}^{(B)}$ converges in probability to $\widehat{\mu}_{rf}^{(\infty)}$. This result is not surprising as almost sure convergence (see (12)) implies convergence in probability. The bound (22) may be used to select the number of trees in practice. For simple random sampling without replacement, the denominator on the right hand-side of (22) can be expressed as

$$\mathbb{P}_\Theta\left(|\widehat{\mu}_{rf,v}^{(B)} - \widehat{\mu}_{rf,v}^{(\infty)}| \geqslant \epsilon\right) \leqslant 2\exp\left(\frac{-B_v\epsilon^2}{2\left(1-\bar{p}_v\right)^2\left(b_y-a_y\right)^2}\right), \tag{23}$$

where $\bar{p}_v := n_{r,v}/n_v$ denotes the response rate. Expression (23) suggests that a larger number of trees would be required for a low response rate $\bar{p}_v$.

7.2 Choice of n_0

The number of observations, n_0, in each terminal node of a tree determines its complexity: a small value of n_0 tends to produce flexible predictions, exhibiting low bias but potentially a large variance. To avoid overfitting and reduce a tree's unnecessary complexity, it is common practice to perform some form of pruning (Hastie et al., 2011). To illustrate the impact of n_0 on the properties of imputed estimators, we conducted a limited simulation study using the same setup as the one described in Section 6. The values of n_0 varied from 1 to $(\mathbb{E}[n_r]+1)/2$, the latter most often leading to a single node in each tree. We computed the

Monte Carlo bias, variance, and mean squared error of the imputed estimator of $\mu_1, \ldots, \mu_5$, the population means of $Y_1, \ldots, Y_5$, respectively. The results are shown in Figure 2.

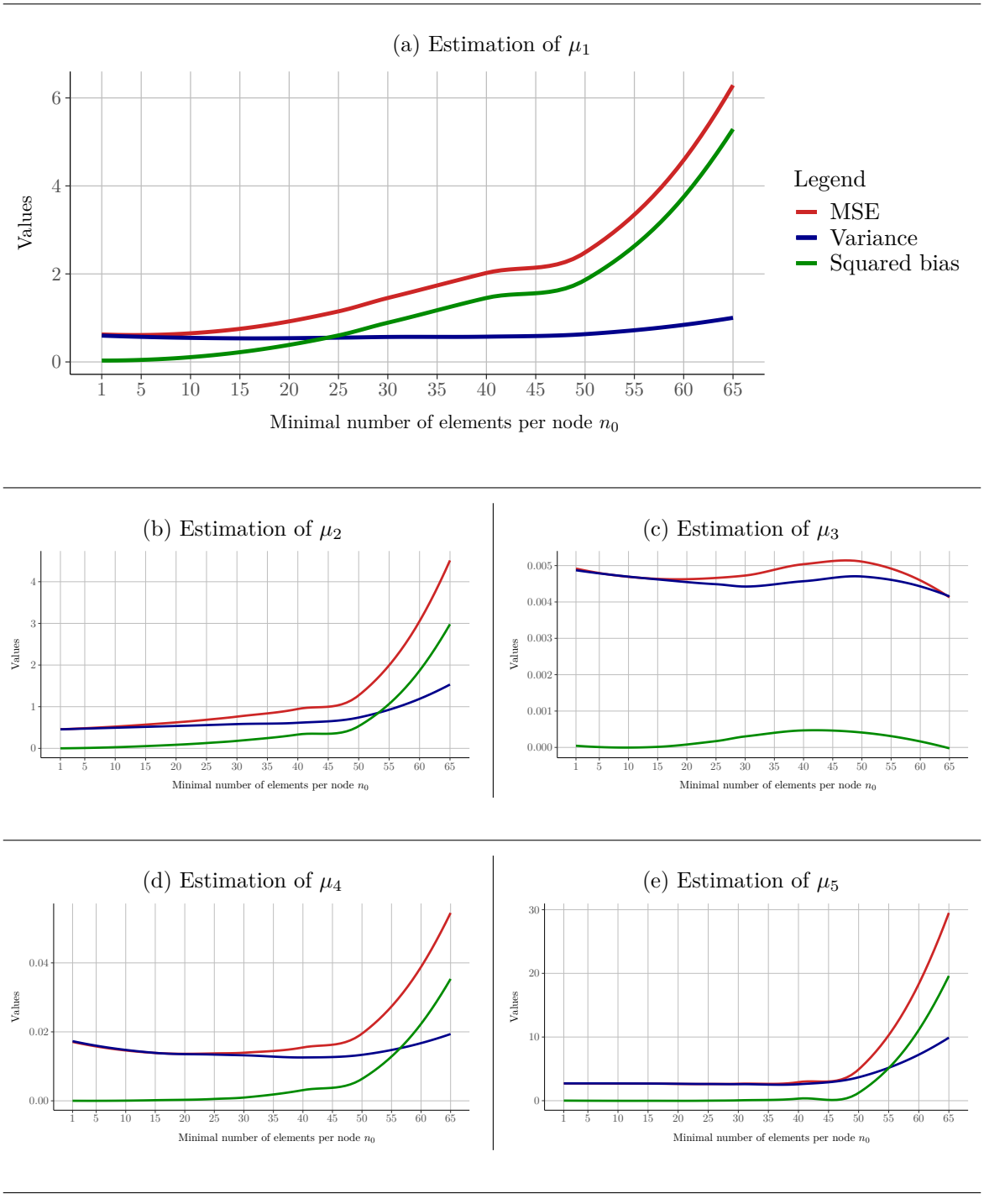

Figure 2: Square bias (green curve), variance (red curve) and mean square error (red curve) of tree imputed estimators as a function of n_0 with $p = 5$ correlated predictors.

From Figure (2), the behavior of the tree imputed estimator was similar across all survey

variables except Y_3. In every scenario, small values of n_0 led to the best results in terms of bias and variance. As n_0 increased, the bias increased. This can be explained by the fact that a large value of n_0 led to shallow trees and somewhat heterogeneous terminal nodes in terms of the survey variable requiring imputation.

For the variable Y_3, both the bias and the variance were essentially identical for all values of n_0. This is an uncommon scenario. Our findings indicate that selecting values for n_0 in the range of 5 to 15 seems to be a safe choice.

7.3 Choice of p_0

In this section, we discuss the choice of p_0, the number of predictors considered at each split.

Proposition 7.2. *Let T_b denote the number of nodes in the b-th tree and let X denote an arbitrary predictor among $X_1, ..., X_{p_v}$. Then,*

$$\mathbb{P}_\Theta\left(\{X \text{ not considered in } \widehat{m}_{rf,v}^{(B)}\}\right) = \prod_{b=1}^{B_v}\left\{1 - \frac{p_{0v}}{p_v}\right\}^{T_b - 1}.$$

Proposition 7.2 suggests that, when the value of p_0 is relatively small compared to p, for a given fixed B, there is a high probability that a predictor X will not be considered. However, in order to achieve an efficient reduction of the potential nonresponse bias, the predictors that are associated with both the survey variable requiring imputation and the response indicators must be included for obtaining the predictions. If p_0 is small compared to p, the predictions will likely fail to incorporate these important predictors. Ultimately, this may result in a biased estimator of the population mean. To cope with this issue, we suggest performing a set of univariate analyses to determine which predictors among the available predictors are related to the response indicator. The selected predictors would then be considered at each split with probability one. For the non-selected predictors, we select, as usual, a subset of predictors at random.

To illustrate the effect of p_0 on the quality of the resulting estimators, we conducted a simulation study using the same setup as the one described in Section 6. Recall from (21) that the predictors X_1-X_5 were related to the response indicators. Again, we were interested in estimating $\mu_1, \mu_2, \ldots, \mu_5$, the population means of the survey variables Y_1-Y_5, respectively. We considered three choices for p_0: $p_0 = \lfloor\sqrt{p}\rfloor$, $p_0 = p$, and $p_0 = \mathcal{M} := \lfloor\sqrt{p-5}\rfloor + \{X_1, \ldots, X_5\}$.

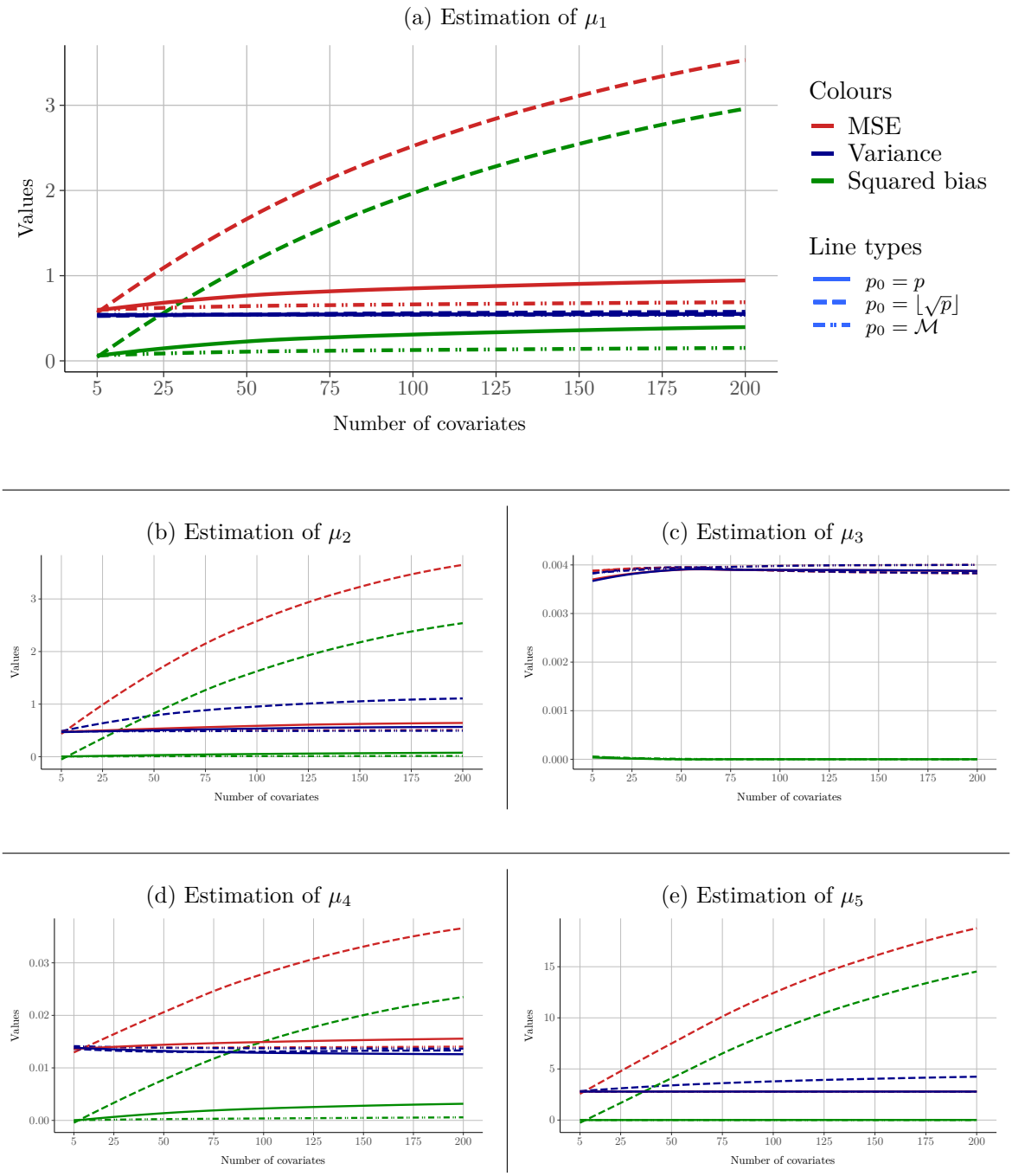

Figure 3: Evolution of the mean squared errors (red curves), squared biases (green curves), and variances (blue curves) of random forests estimators as the number of predictors p increases. Dotted lines represent the choice $p_0 = \lfloor\sqrt{p}\rfloor$, full lines represent $p_0 = p$, combination of both indicates the choice $p_0 = \mathcal{M} := \lfloor\sqrt{p}\rfloor + \text{MAR}$ variables.

For the latter choice, the predictors X_1-X_5 were considered at each split with probability one, as they were associated with the response indicator, while the $p-5$ remaining predictors

were subject to a random selection. In our experiments, the number of predictors p ranged between 5 and 200.

Results are shown in Figure (3). We start by noting that the default choice in most software packages, $p_0 = \lfloor \sqrt{p} \rfloor$, produced biased estimators, in general. The bias increased as the number of predictors p increased. This can be explained by the fact that, as the number of predictors increased, the likelihood of the predictors X_1 through X_5 being considered at each split diminished significantly. Therefore, for a large number of predictors, it is likely that a significant proportion of the predictions will not incorporate X_1-X_5. The choice $p_0 = p$ led to good results in all the scenarios, as expected. Finally, our proposal, $p_0 = \mathcal{M}$, led to results as good or better than the ones obtained with $p_0 = p$. Note that this choice led to good results even when p was larger than the number of respondents n_r. For these reasons, we recommend choosing $p_0 = \mathcal{M}$ in practice.

8 Final remarks

In this article, we have studied the theoretical properties of the regression tree and random forest imputed estimators. In particular, we have established the mean square consistency of the imputed estimator in a high-dimensional setting which allowed the number of predictors to diverge. In addition, we proposed a novel variance estimator based on a K-fold cross-validation procedure. Unlike the customary variance estimator based on a first-order Taylor expansion, our simulation results suggest that the proposed variance estimator performs well in terms of bias and coverage rate of normal-based confidence intervals.

This work constitutes a first step toward understanding the behavior of regression trees and random forest imputed estimators with survey data. Several questions remain open. First, establishing the convergence rate and asymptotic behavior of these imputed estimators would be desirable. This would require techniques different from those used in this paper, as minimax rates can be much slower in the case of nonparametric methods than the actual convergence rate of imputed estimators. Establishing a central limit theorem would also be useful. The proposed variance estimator based on a K-fold cross-validation shares some common features with the so-called cross-fitting variance estimator that has been studied in the context of causal inference for estimating the average treatment effect; e.g., Wager

et al. (2016). Cross-fitting is a technique used to reduce overfitting and bias of the point and variance estimators. It involves splitting the dataset into multiple folds. For each fold, part of the data is used to estimate nuisance parameters, while the other part is used to estimate the point and variance estimator. We refer the reader to the seminal works of Chernozhukov et al. (2017); Newey and Robins (2018) and Smucler et al. (2019) for additional details on sample splitting procedures. Additionally, employing cross-fitting procedures simplifies proving the consistency and asymptotic normality of the point estimators. The application of cross-fitting procedures to both point and variance estimation procedures in the context of survey data is currently under investigation.

Moreover, other useful tools could include the theory of incomplete U-statistics that were used for the estimator of the regression function estimator (Mentch and Hooker, 2016; Zhou et al., 2019; Xu et al., 2024).

Appendix

8.1 Supplementary simulation results: Poisson sampling

We present the results of a simulation study that assesses the performance of point and variance estimators in the context of Poisson sampling. The simulation setup used in this section is identical to the one used in Section 6, except that simple random sampling without replacement was replaced by Poisson sampling, whereby the first-order inclusion probabilities were defined as

$$\pi_k := \frac{n x_{k1}^2}{\sum_{l \in U} x_{l1}^2}, \qquad k \in U.$$

This ensured that all inclusion probabilities were strictly positive, with values ranging from 0.004 to 0.15. Table 10 presents the correlation between the π_k's and the survey variables $Y_1, ..., Y_5$, for the cases of both independent and correlated covariates. As shown in Table 10,

	Y_1	Y_2	Y_3	Y_4	Y_5
Independent covariates	0.27	0.71	-0.17	0.24	0.06
Correlated covariates	0.72	0.72	-0.01	0.47	0.52

Table 10: Correlation between the survey variables and the inclusion probabilities.

the first-order inclusion probabilities were highly correlated to some of the survey variables, reaching correlations up to 0.72. The results pertaining to the behavior of point estimators

are presented in Tables 11-14. The results of the behavior of variance estimators based on regression trees, are presented in Table 15 for the case of independent covariates. As the results for dependent covariates are nearly identical to those for independent covariates, we have chosen to omit their presentation.

Overall, for both point and variance estimation, the conclusions of the simulation results are very similar to those presented in Section 6 in the case of simple random sampling without replacement. The only slight difference arises in variance estimation: both the naïve variance estimators and those based on a first-order Taylor expansion showed less bias compared to their counterparts in simple random sampling without replacement. This can be attributed to the relatively high dispersion of the sampling weights π_k^{-1}, which makes the complete data estimator unstable. Consequently, the sampling variance contributes significantly to the total variance, reducing the relative impact of nonresponse on the variance estimators. Nonetheless, the variance estimator based on a K-fold cross-validation procedure behaved very well, and improved on the variance estimators based on a first-order Taylor expansion.

Survey variable	MC measure	Imputed estimators				
		LR	NN	KNN	CART	RF
Y_1	RB	0.0	0.7	1.2	1.3	0.8
	RE	101.7	106.1	108.3	109.5	105.3
Y_2	RB	-0.1	0.3	0.9	11.6	1.6
	RE	921.8	107.9	107.5	309.3	109.9
Y_3	RB	0.0	-0.8	-0.4	0.6	-0.4
	RE	137.7	119.1	111.7	124.6	113.8
Y_4	RB	0.0	0.8	1.6	1.6	0.9
	RE	148.6	157.2	145.4	147.2	140.8
Y_5	RB	0.0	-0.7	-0.6	4.0	0.3
	RE	345.6	114.6	113.1	126.3	101.9

Table 11: Monte Carlo Simulation Results for $p = 5$ and correlated predictors with Poisson sampling.

8.2 Proofs and technical details

We will start by establishing a preliminary result that will prove useful in establishing the mean square consistency of imputed estimators obtained through regression trees and random

Survey variable	MC measure	Imputed estimators				
		LR	NN	KNN	CART	RF
Y_1	RB	-0.1	0.6	0.9	0.8	0.6
	RE	101.2	104.4	105.0	105.1	103.6
Y_2	RB	-0.3	0.9	2.3	5.4	2.1
	RE	183.0	124.4	127.6	193.9	128.2
Y_3	RB	-0.1	0.8	0.5	-0.7	-0.3
	RE	127.4	119.9	118.9	121.4	114.1
Y_4	RB	-0.1	0.5	0.7	1.2	0.8
	RE	131.0	138.5	127.0	132.0	125.2
Y_5	RB	-0.2	-2.0	-2.2	1.8	0.5
	RE	147.2	133.8	128.1	105.8	102.6

Table 12: Monte Carlo Simulation Results for $p = 5$ and independent predictors with Poisson sampling.

Survey variable	MC measure	Imputed estimators				
		LR	NN	KNN	CART	RF
Y_1	RB	-0.1	1.1	1.5	1.4	1.3
	RE	108.2	110.6	111.9	110.4	108.8
Y_2	RB	-0.2	3.6	4.3	14.2	3.0
	RE	3742.2	136.6	135.0	426.0	118.6
Y_3	RB	-0.1	-0.4	-0.3	0.8	0.1
	RE	202.7	132.4	117.8	131.8	115.1
Y_4	RB	-0.7	1.7	2.4	1.7	1.6
	RE	253.7	175.2	155.5	155.8	141.8
Y_5	RB	-10.5	3.0	3.1	3.6	0.4
	RE	550.6	164.5	146.7	121.8	101.9

Table 13: Monte Carlo Simulation Results for $p = 90$ and correlated predictors with Poisson sampling.

forests.

Result 8.1. *We assume that (H1) holds. Let* $\{\widetilde{m}_v\}_{v\in\mathbb{N}}$ *be a sequence of regression function estimates fitted on* $D_{U_v} := \{(\mathbf{x}_k, y_k) \ ; \ k \in U_v\}$ *and let* $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}$ *be independent of* D_{U_v}*. Let* $\{\widehat{m}_v\}_{v\in\mathbb{N}}$ *be the corresponding estimates fitted on* $D_{r_v} = \{(\mathbf{x}_k, y_k) \ ; \ k \in S_{r,v}\}$*. If:*

i) The sequence of population predictors $\{\widetilde{m}_v\}_{v\in\mathbb{N}}$ *satisfies*

Survey variable	MC measure	Imputed estimators				
		LR	NN	KNN	CART	RF
Y_1	RB	0.0	2.2	2.3	1.4	1.6
	RE	104.3	120.0	119.5	109.0	110.0
Y_2	RB	-0.7	18.7	21.5	5.5	6.3
	RE	367.0	860.2	931.2	184.8	180.3
Y_3	RB	0.9	-2.6	-2.9	-2.2	-2.8
	RE	170.7	148.4	135.6	131.6	122.9
Y_4	RB	0.4	3.3	3.4	2.6	2.7
	RE	211.3	190.4	168.3	144.1	135.9
Y_5	RB	-1.2	8.6	9.4	1.8	0.5
	RE	262.6	324.7	282.6	105.8	102.4

Table 14: Monte Carlo Simulation Results for $p = 90$ and independent predictors with Poisson sampling.

Survey variable	Estimator	$n_0 = 5$		$n_0 = 10$		$n_0 = 20$	
		RB	Coverage	RB	Coverage	RB	Coverage
Y_1	Naïve	-1.6	94.7	1.0	95.0	-1.0	94.9
	Shao	-1.1	94.7	1.4	95.0	-1.0	94.9
	CV	-0.9	94.7	1.0	95.0	-1.7	94.8
Y_2	Naïve	-23.5	91.6	-27.2	90.1	-33.3	85.5
	Shao	-13.0	93.5	-9.7	93.2	-8.2	91.6
	CV	1.8	95.2	2.0	94.9	0.2	93.1
Y_3	Naïve	-19.1	91.6	-19.6	91.3	-23.3	90.3
	Shao	-9.9	93.1	-5.5	93.7	-5.9	93.3
	CV	5.2	95.0	4.3	94.8	0.1	93.9
Y_4	Naïve	-26.7	90.5	-23.6	91.5	-23.4	90.9
	Shao	-15.5	92.7	-8.0	94.0	-5.8	94.1
	CV	-0.4	94.8	-0.3	94.9	-1.7	94.6
Y_5	Naïve	-3.1	93.7	0.1	94.2	-1.5	94.2
	Shao	-3.1	93.7	-0.2	94.2	0.4	94.8
	CV	-1.2	94.0	1.4	94.5	5.2	95.3

Table 15: Monte-Carlo simulation results for various variance estimators across different sample sizes (n_0) with $p = 5$ independent covariates with Poisson sampling.

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\widetilde{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2\right] = 0,$$

with a convergence rate denoted by γ_v*, with* $\gamma_v \to 0$.

ii) There exists a positive constant C*, independent of* v*, such that*

$$\mathbb{E}\left\{\left(\widehat{m}_v(\mathbf{x}) - m(\mathbf{x})\right)^2 \;\middle|\; \boldsymbol{r}_v, \boldsymbol{X}_v, \boldsymbol{I}_v\right\} \leqslant C. \qquad a.s.$$

Then, the sequence of imputed estimators $\{\widehat{\mu}_{\widehat{m}_v}\}_{v\in\mathbb{N}}$ *satisfies*

$$\mathbb{E}\left(\widehat{\mu}_{\widehat{m}_v} - \mu_v\right)^2 = \mathcal{O}(\max(\gamma_v, 1/n_v)). \tag{24}$$

Condition (i) is satisfied for a large number of (parametric and nonparametric) estimators of the regression function, including, for instance, k-nearest neighbors and kernel regression, among others; see Györfi et al. (2006). Result 8.1 suggests that, in order to build a consistent imputed estimator, it is enough to use a consistent regression function estimator to impute the missing values. Note that Result 8.1 holds in a high-dimensional setting in which the number of predictors $\{p_v\}_{v\in\mathbb{N}}$ is allowed to increase to infinity, provided that conditions i) and ii) of Result 8.1 are satisfied.

Proof. We write

$$\mathbb{E}\left[\left(\widehat{\mu}_{\widehat{m}_v} - \mu_v\right)^2\right] \leqslant 2\mathbb{E}\left[\left(\widehat{\mu}_{\widehat{m}_v} - \widehat{\mu}_{\pi,v}\right)^2\right] + 2\mathbb{E}\left[\left(\widehat{\mu}_{\pi,v} - \mu_v\right)^2\right], \tag{25}$$

where $\widehat{\mu}_{\pi,v}$ denotes the complete data estimator given by (1). The second term of the right hand-side of (25) is the mean squared error of $\widehat{\mu}_{\pi,v}$ and it can be shown that under the assumption (H1), $\mathbb{E}\left[\left(\widehat{\mu}_{\pi,v} - \mu_v\right)^2\right] = \mathcal{O}(n_v^{-1})$ (Robinson and Särndal, 1983; Breidt and Opsomer, 2000). Consider now the first term on the right hand-side of (25), which can be written as follows:

$$\widehat{\mu}_{\widehat{m}_v} - \widehat{\mu}_{\pi,v} = \frac{1}{N_v}\sum_{k\in S_v}\left\{\frac{(1-r_k)}{\pi_k}\left(\widehat{m}_v(\mathbf{x}_k) - y_k\right)\right\}.$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{\mu}_{\widehat{m}_v} - \widehat{\mu}_{\pi,v}\right)^2\right] \leqslant\; & 2\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S_v}\frac{(1-r_k)}{\pi_k}\cdot\left\{\widehat{m}_v(\mathbf{x}_k) - m(\mathbf{x}_k)\right\}\right)^2\right] \\ & + 2\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S_v}\frac{(1-r_k)}{\pi_k}\left(m(\mathbf{x}_k) - y_k\right)\right)^2\right].\end{aligned} \tag{26}$$

Write

$$\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S_v}(1-r_k)\left(m(\mathbf{x}_k)-y_k\right)\right)^2\right] = \mathbb{E}\left[\frac{1}{N_v^2}\sum_{\substack{k,\ell\in S_v\\ \ell\neq k}}\frac{(1-r_k)}{\pi_k}\frac{(1-r_\ell)}{\pi_l}\times\epsilon_k\epsilon_\ell\right] + \mathbb{E}\left[\frac{1}{N_v^2}\sum_{k\in S_v}\left(\frac{(1-r_k)}{\pi_k}\right)^2\epsilon_k^2\right]. \tag{27}$$

For the first term on the right hand-side of (27), we use the law of total expectation to obtain

$$\mathbb{E}\left[\frac{1}{N_v^2}\sum_{\substack{k,\ell\in S_v\\ l\neq k}}\frac{(1-r_k)}{\pi_k}\frac{(1-r_l)}{\pi_l}\times\epsilon_k\epsilon_\ell\right] = \mathbb{E}\left[\frac{1}{N_v^2}\sum_{\substack{k,\ell\in S_v\\ \ell\neq k}}\frac{(1-r_k)}{\pi_k}\frac{(1-r_\ell)}{\pi_\ell}\mathbb{E}\left[\epsilon_k\epsilon_\ell\middle|\boldsymbol{X}_v,\boldsymbol{I}_v,\boldsymbol{r}_v\right]\right] = 0,$$

since the random variables ϵ_k and ϵ_ℓ are independent for all $k\neq\ell$ and $\mathbb{E}\left[\epsilon_k|\mathbf{x}_k\right]=0$. For the second term on the right hand-side of (27), we have that

$$\mathbb{E}\left[\frac{1}{N_v^2}\sum_{k\in S_v}\left(\frac{(1-r_k)}{\pi_k}\right)^2\epsilon_k^2\right] \leqslant \frac{N_v}{\lambda^2 N_v^2}\max_{k\in U_v}\mathbb{E}\left[\epsilon_k^2\right] = \frac{\sigma^2}{\lambda^2 N_v} = \mathcal{O}(N_v^{-1}).$$

It remains to bound the first term on the right hand-side of (26). Bounding arguments ensure that

$$\mathbb{E}\left[\left(\frac{1}{N_v}\sum_{k\in S_v}\frac{(1-r_k)}{\pi_k}\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)\right)^2\right] \leqslant \frac{n_v}{\lambda^2 N_v}\mathbb{E}\left[\frac{1}{N_v}\sum_{k\in S_{m,v}}\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2\right].$$

Now, Condition ii) implies that there exists a positive constant $C>0$, independent of v, such that

$$\mathbb{E}\left[\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2\middle|\boldsymbol{r}_v,\boldsymbol{X}_v,\boldsymbol{I}_v\right]\leqslant C, \qquad \text{a.s.}$$

From Condition ii) and the assumptions MAR and that the sampling design is non-informative, it follows that, uniformly,

$$\mathbb{E}\left[\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2\middle|\boldsymbol{r}_v,\boldsymbol{X}_v,\boldsymbol{I}_v\right]\xrightarrow{\mathbb{P}}0.$$

Hence, by the Lebesgues dominated convergence theorem,

$$\lim_{v\to\infty}\mathbb{E}\left[\frac{1}{N_v}\sum_{k\in S_{m,v}}\mathbb{E}\left[\left(\widehat{m}_v(\mathbf{x}_k)-m(\mathbf{x}_k)\right)^2\middle|\boldsymbol{r}_v,\boldsymbol{X}_v,\boldsymbol{I}_v\right]\right]=0,$$

with the rate $O(\gamma_v)$. The result follows. ■

Proof of Result 4.1. We begin by noting that, from Corollary 4.3 of Klusowski and Tian (2024), it follows that the sequence $\{\widetilde{m}_{tree,v}\}_{v\in\mathbb{N}}$ of tree predictors fitted on D_{N_v} is consistent in L^2 for m, meaning

$$\lim_{v\to\infty}\mathbb{E}\left[\left(\widetilde{m}_{tree,v}(\mathbf{x})-m_v(\mathbf{x})\right)^2\right]=0,$$

which is Condition i) of Result 8.1. Since Y is assumed to be almost surely bounded, it follows that there exists $C>0$, satisfying

$$\mathbb{E}\left\{\left(\widehat{m}_{tree,v}(\mathbf{x})-m_v(\mathbf{x})\right)^2\ \middle|\ \boldsymbol{r}_v,\boldsymbol{X}_v,\boldsymbol{I}_v\right\}\leqslant C. \qquad \text{a.s.}$$

Therefore, Condition ii) of Result 8.1 holds as well. Hence, Result 8.1 ensures the mean-square consistency of $\{\widehat{\mu}_{tree,v}\}_{v\in\mathbb{N}}$. ■

Proof of Proposition 4.1.

1. We can write

$$\begin{aligned}\mathbb{E}(\widehat{\mu}_{rf,v}^{(B)}-\mu_v)^2&=\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B}(\widehat{\mu}_{tree,v}^{(b)}-\mu_v)\right]^2\\&\leqslant\frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\left[\mathbb{E}_\Theta\left(\widehat{\mu}_{tree,v}^{(b)}-\mu_v\right)^2\right]=\mathbb{E}\left[\mathbb{E}_\Theta\left(\widehat{\mu}_{tree}^{(b)}-\mu_v\right)^2\right]=\mathbb{E}(\widehat{\mu}_{tree}^{(b)}-\mu_v)^2,\end{aligned}$$

   using the fact that $\{\Theta_b\}_{b=1,\ldots,B}$ are i.i.d.; moreover, equality holds if and only if there exists C such that $\widehat{\mu}_{tree}^{(b)}=C$, almost surely, for all $b=1,\ldots,B_v$, which implies that $\widehat{\mu}_{rf,v}^{(B)}=C$ almost surely, meaning that the forest is degenerate.

2. The proof essentially follows ideas described in Scornet (2016). Write

$$\begin{aligned}\left(\widehat{\mu}_{rf,v}^{(B)}-\mu_v\right)^2&=\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}+\widehat{\mu}_{rf,v}^{(\infty)}-\mu_v\right)^2\\&=\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}\right)^2+\left(\widehat{\mu}_{rf,v}^{(\infty)}-\mu_v\right)^2+2\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}\right)\left(\widehat{\mu}_{rf,v}^{(\infty)}-\mu_v\right).\end{aligned}\tag{28}$$

   Next, note that

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}\right)\left(\widehat{\mu}_{rf,v}^{(\infty)}-\mu_v\right)\right]=\mathbb{E}\left[\mathbb{E}_\Theta\left[\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}\right)\right]\left(\widehat{\mu}_{rf,v}^{(\infty)}-\mu_v\right)\right]=0.$$

   Taking expectations on both sides of (28) leads to

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(B)}-\mu_v\right)^2\right]=\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}\right)^2\right]+\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(\infty)}-\mu_v\right)^2\right],$$

so that

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(B)}-\mu\right)^2\right]-\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(\infty)}-\mu_v\right)^2\right]=\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}\right)^2\right]\geqslant 0.$$

3. Write

$$\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}=\frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}_{rf,v}^{(B)}(\mathbf{x}_k)-\widehat{m}_{rf,v}^{(\infty)}(\mathbf{x}_k)}{\pi_k},$$

so that

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}\right)^2\right]&=\frac{1}{N_v^2}\cdot\mathbb{E}\left[\left(\sum_{k\in S_m}\frac{\widehat{m}_{rf,v}^{(B)}(\mathbf{x}_k)-\widehat{m}_{rf,v}^{(\infty)}(\mathbf{x}_k)}{\pi_k}\right)^2\right]\\&\leqslant\frac{n_v}{N_v^2}\cdot\mathbb{E}\left[\sum_{k\in S_m}\frac{\left(\widehat{m}_{rf,v}^{(B)}(\mathbf{x}_k)-\widehat{m}_{rf,v}^{(\infty)}(\mathbf{x}_k)\right)^2}{\pi_k^2}\right]\\&\leqslant\frac{n_vN_v}{N_v^2\lambda^2}\cdot\max_{k\in U_v}\mathbb{E}\left[\left(\widehat{m}_{rf,v}^{(B)}(\mathbf{x}_k)-\widehat{m}_{rf,v}^{(\infty)}(\mathbf{x}_k)\right)^2\right]\end{aligned}$$

Now, using Theorem 3.3 of Scornet (2016), there exists a positive constant C such that, uniformly,

$$\mathbb{E}\left[\left(\widehat{m}_{rf,v}^{(B)}(\mathbf{x}_k)-\widehat{m}_{rf,v}^{(\infty)}(\mathbf{x}_k)\right)^2\right]\leqslant\frac{C}{B_v},$$

leading to

$$\mathbb{E}\left[\left(\widehat{\mu}_{rf,v}^{(B)}-\widehat{\mu}_{rf,v}^{(\infty)}\right)^2\right]\leqslant\frac{Cn_vN_v}{N_v^2\lambda^2B_v}=\mathcal{O}\left(\frac{1}{B_v}\right).$$

■

Proof of Result 4.2. Corollary 1 of Scornet (2016) leads to the consistency of a sequence infinite forest estimators $\{\widehat{m}_{urf,v}^{(\infty)}\}_{v\in\mathbb{N}}$ in a framework in which the dimension p is fixed. We extend their proof to a high-dimensional asymptotic framework. To that aim, we use Stone's Theorem, see e.g., Györfi et al. (2006), page 56. We begin by noting that, in our framework, verifying Stone's theorem conditions is enough to ensure consistency. That is, as shown by Biau et al. (2008) and Scornet (2016), we must prove that

$$\text{for all } K>0,\quad \lim_{v\to\infty}\mathbb{P}\left\{\text{Card}\left(A_v\left(\mathbf{x},\Theta\right)\right)>K\right\}=1, \tag{29}$$

$$\text{for all } \epsilon>0,\quad \lim_{v\to\infty}\mathbb{P}\left\{\text{diam}\left(A_v\left(\mathbf{x},\Theta\right)\right)>\epsilon\right\}=0, \tag{30}$$

where $\text{diam}\left(A_v\left(\mathbf{x},\Theta\right)\right)$ is used to denote the diameter of the hyper-rectangle $A_v\left(\mathbf{x},\Theta\right)$, i.e., the maximal distance between two points in the rectangle. The proof of (29) given by Scornet (2016)

continues to hold in a high-dimensional asymptotic framework. It is thus enough to prove (30). To that aim, let $d_{v,j}$ denote the length of the j-th side of the rectangle containing $\mathbf{x}$, and $\mathbf{d} := [d_{1,v}, d_{2,v}, ..., d_{p_v,v}]^\top$ with $d_{j,v} > 0$ for all $j = 1, \dots p_v$. Let $\epsilon > 0$ and write

$$\mathbb{P}\left(\text{diam}\left(A_v\left(\mathbf{x}, \Theta\right)\right) > \epsilon\right) \leqslant \mathbb{P}\left(||\mathbf{d}||_2 > \epsilon\right) \leqslant \mathbb{P}\left(||\mathbf{d}||_1 > \epsilon\right) \leqslant \frac{\mathbb{E}\left[\sum_{j=1}^{p_v} d_{j,v}\right]}{\epsilon} = p_v \frac{\mathbb{E}\left[d_{1,v}\right]}{\epsilon},$$

using norms' inequality, Markov's inequality, and symmetry of the dimensions. Let $K_{1,v}$ denote the number of times the leaf containing $\mathbf{x}$ has been cut along the first coordinate. Then, as noted by Biau et al. (2008), we can use the following inequality,

$$\mathbb{E}\left[d_{1,v}\right] \leqslant \mathbb{E}\left[\left(\frac{3}{4}\right)^{K_{1,v}}\right]$$

to obtain that

$$\mathbb{E}\left[d_{1,v}\right] \leqslant \sum_{l=0}^{K_v} \binom{K_v}{l} \left(\frac{3}{4}\right)^l \left(\frac{1}{p_v}\right)^l \left(1 - \frac{1}{p_v}\right)^{K_v - l} = \left(1 - \frac{1}{4p_v}\right)^{K_v}.$$

Therefore, combining these inequalities leads to

$$\mathbb{P}(\text{diam}\left(A_v\left(\mathbf{x}, \Theta\right)\right) > \epsilon) \leqslant \frac{p_v}{\epsilon}\left(1 - \frac{1}{4p_v}\right)^{K_v}.$$

Hence, under our conditions, both (29) and (30) hold, leading to

$$\lim_{v \to \infty} \mathbb{E}\left[\left\{\widehat{m}_v^{(\infty)}(\mathbf{x}) - m_v(\mathbf{x})\right\}^2\right] = 0.$$

Applying Result 8.1 gives the consistency of the infinite uniform forest estimator. Moreover, from Proposition 4.1, we have

$$0 \leqslant \mathbb{E}\left[\left(\widehat{\mu}_{urf,v}^{(B)} - \mu_v\right)^2\right] - \mathbb{E}\left[\left(\widehat{\mu}_{urf,v}^{(\infty)} - \mu_v\right)^2\right] \leqslant \frac{C}{B_v}.$$

Thus, if we consider large forests (i.e., with an increasing number of trees), the sequences $\mathbb{E}\left[\left(\widehat{\mu}_{urf,v}^{(B)} - \mu_v\right)^2\right]$ and $\mathbb{E}\left[\left(\widehat{\mu}_{urf,v}^{(\infty)} - \mu_v\right)^2\right]$ must have the same limit. Hence, $\lim_{v\to\infty} \mathbb{E}\left[\left(\widehat{\mu}_{urf,v}^{(B)} - \mu_v\right)^2\right] = 0$, which concludes the proof. ■

Proof of Result 4.3. It follows from (Klusowski and Tian, 2024, Corollary 7.2) that, under our conditions, the sequence $\{\widetilde{m}_{brf,v}^{(B)}\}_{v\in\mathbb{N}}$ of tree predictors fitted on D_{N_v} is consistent in L^2

for m, meaning

$$\lim_{v\to\infty} \mathbb{E}\left[\left(\widetilde{m}^{(\infty)}_{brf,v}(\mathbf{x}) - m_v(\mathbf{x})\right)^2\right] = 0,$$

which is Condition i) of Result 8.1. Clearly, Condition ii) of Result 8.1 holds also in this framework so that Result 8.1 gives the consistency of the imputed estimator based on the infinite Breiman's random forest. Consistency of the finite forest imputed estimator follows from applying Proposition 4.1. ■

Proof of Proposition 5.1. By the law of iterated variance and relation (12),

$$\mathbb{V}\left(\widehat{\mu}^{(B)}_{rf,v} - \mu_v\right) = \mathbb{V}\left(\mathbb{E}_\Theta\left[\widehat{\mu}^{(B)}_{rf,v} - \mu_v\right]\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}^{(B)}_{rf,v} - \mu_v\right)\right] = \mathbb{V}\left(\widehat{\mu}^{(\infty)}_{rf,v} - \mu_v\right) + \mathbb{E}\left[\mathbb{V}_\Theta\left(\widehat{\mu}^{(B)}_{rf,v} - \mu_v\right)\right].$$

Relation (14) is proved. Next, we have

$$\mathbb{V}_\Theta\left(\widehat{\mu}^{(B)}_{rf,v} - \mu_v\right) = \mathbb{V}_\Theta\left(\widehat{\mu}^{(B)}_{rf,v}\right) = \mathbb{V}_\Theta\left(\frac{1}{B_v}\sum_{b=1}^{B_v}\widehat{\mu}^{(b)}_{tree,v}\right) \overset{(4)}{=} \frac{1}{B_v}\cdot\mathbb{V}_\Theta\left(\widehat{\mu}^{(1)}_{tree,v}\right),$$

where equality (4) follows from the fact that, as detailed in the proof of Proposition 7.1, conditionally on everything but $\{\Theta_b\}_{b=1}^{B_v}$, $\{\widehat{\mu}^{(b)}_{tree,v}\}_{b=1}^{B_v}$ is a sequence of i.i.d. random variables. Now, for any $b \in \{1, 2, ..., B_v\}$,

$$\mathbb{V}_\Theta\left(\widehat{\mu}^{(b)}_{tree,v}\right) = \mathbb{V}_\Theta\left(\frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}^{(b)}_{tree,v}(\mathbf{x}_k)}{\pi_k}\right) \leqslant \mathbb{E}_\Theta\left[\left(\frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}^{(b)}_{tree,v}(\mathbf{x}_k)}{\pi_k}\right)^2\right] \leqslant \frac{n_v^2}{N_v^2}\left(\max\{|a_y|, |b_y|\}\max_{k\in U} d_k\right)^2.$$

This concludes the proof. ■

Proof of Proposition 7.1. Observe that

$$\widehat{\mu}^{(B)}_{rf,v} - \widehat{\mu}^{(\infty)}_{rf,v} = \frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}^{(B)}_{rf,v}(\mathbf{x}_k) - \widehat{m}^{(\infty)}_{rf,v}(\mathbf{x}_k)}{\pi_k},$$

so that, for $\epsilon > 0$,

$$\mathbb{P}_\Theta\left(|\widehat{\mu}^{(B)}_{rf,v} - \widehat{\mu}^{(\infty)}_{rf,v}| \geqslant \epsilon\right) = \mathbb{P}_\Theta\left(\left|\frac{1}{B_v}\sum_{b=1}^{B_v}\left\{\frac{1}{N_v}\sum_{k\in S_m}\frac{\widehat{m}^{(b)}_{tree,v}(\mathbf{x}_k) - \widehat{m}^{(\infty)}_{rf,v}(\mathbf{x}_k)}{\pi_k}\right\}\right| \geqslant \epsilon\right).$$

Define $\widehat{d}^{\,(b)} := \frac{1}{N_v}\sum_{k\in S_m}\pi_k^{-1}\left(\widehat{m}^{(b)}_{tree,v}(\mathbf{x}_k) - \widehat{m}^{(\infty)}_{rf,v}(\mathbf{x}_k)\right)$. Note that, given the predictors, the sample membership indicators, the survey variable, and the nonresponse indicators, the sequence $\{\widehat{m}^{(b)}_{tree,v}\}_{b=1}^{B_v}$ is a sequence of independently and identically distributed (according to $\mathbb{P}_\Theta$) random variables. The same holds therefore for the sequence $\{\widehat{d}^{\,(b)}\}_{b=1}^{B_v}$. Moreover, in our framework, these

are zero mean bounded random variables. To see that, recall that there exists $a_y < b_y$ such that $Y \in [a_y \ ; \ b_y]$, almost surely. Hence, for all $b \in \{1, 2, ..., B_v\}$ and $k \in S_m$,

$$a_y - b_y \leqslant \widehat{m}^{(b)}_{tree,v}(\mathbf{x}_k) - \widehat{m}^{(\infty)}_{rf,v}(\mathbf{x}_k) \leqslant b_y - a_y. \qquad a.s.$$

Therefore,

$$\frac{n_{m,v}}{N_v} \cdot \frac{a_y - b_y}{\min_{k \in U} \pi_k} \leqslant \widehat{d}^{\,(b)} \leqslant \frac{n_{m,v}}{N_v} \cdot \frac{b_y - a_y}{\min_{k \in U_v} \pi_k}, \qquad a.s.$$

Thus, for $\epsilon > 0$,

$$\mathbb{P}_\Theta \left(|\widehat{\mu}^{(B)}_{rf,v} - \widehat{\mu}^{(\infty)}_{rf,v}| \geqslant \epsilon \right) = \mathbb{P}_\Theta \left(\frac{1}{B_v} \left| \sum_{b=1}^{B_v} \widehat{d}^{\,(b)} \right| \geqslant \epsilon \right) \overset{(3)}{\leqslant} 2 \exp \left(\frac{-2B_v\epsilon^2}{4\frac{n^2_{m,v}}{N_v^2} \left(\frac{b_y - a_y}{\min_{k \in U} \pi_k} \right)^2} \right),$$

where (3) follows from Hoeffding inequality for bounded random variables and the fact that $\widehat{d}^{\,(b)}, b = 1, \ldots, B_v$ are i.i.d. random variables with $\mathbb{E}_\Theta(\widehat{d}^{\,(b)}) = 0$. ■

Proof of Proposition 7.2. Let X be an arbitrary covariate among $X_1, X_2, ..., X_p$. Denote by $\mathcal{S}$ the set of predictors considered (at least once) for splitting in $\widehat{m}^{(B)}_{rf,v}$, and $\mathcal{S}_b$ those considered in $\widehat{m}^{(b)}_{tree,v}$. Basic graph theory reveals that, if T_b denotes the number of terminal nodes of the b-th tree $\widehat{m}^{(b)}_{tree,v}$, then the number of splits in $\widehat{m}^{(b)}_{tree}$ is $T_b - 1$. Finally, let $P_{b,j}$ denote the set of predictors considered for splitting in the j-th split $\widehat{m}^{(b)}_{tree,v}$. We may then write

$$\mathbb{P}\left(X \notin \mathcal{S}\right) = \mathbb{P}\left(\bigcap_{b=1}^{B_v} \left(X \notin \mathcal{S}_b\right) \right) = \mathbb{P}\left(\bigcap_{b=1}^{B_v} \bigcap_{j=1}^{T_b - 1} \left(X \notin P_{b,j}\right) \right) = \prod_{b=1}^{B_v} \left(1 - \frac{p_0}{p}\right)^{T_b - 1}$$

by independence between each draw of predictors.

■

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095.

Biau, G., Devroye, L., and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(Sep):2015–2033.

Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1023–1053.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The annals of Statistics*, 30(4):927–961.

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, T. and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265.

Chi, C.-M., Vossler, P., Fan, Y., and Lv, J. (2022). Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438.

Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Dagdoug, M., Goga, C., and Haziza, D. (2023a). Imputation Procedures in Surveys Using Nonparametric and Machine Learning Methods: an Empirical Comparison. *Journal of Survey Statistics and Methodology*, 11(1):141–188.

Dagdoug, M., Goga, C., and Haziza, D. (2023b). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542):1234–1251.

De Moliner, A. and Goga, C. (2018). Sample-based estimation of mean electricity consumption curves for small domains. *Survey Methodology*, 44(2):193–214.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Díaz-Uriarte, R. and de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):3.

Earp, M., Toth, D., Phipps, P., and Oslund, C. (2018). Assessing nonresponse in a longitudinal establishment survey using regression trees. *Journal of Official Statistics*, 34(2):463–481.

Fay, R. (1991). *A design-based perspective on missing data variance*. US Census Bureau.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., and Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, 108(1):10–19.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on barro colorado island — digital soil mapping using random forests analysis. *Geoderma*, 146(1-2):102–113.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hamza, M. and Larocque, D. (2005). An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*, 75(8):629–643.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C., editors, *Handbook of statistics*, volume 29A, pages 215–246. Elsevier.

Haziza, D. and Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: a critical review. *Japanese Journal of Statistics and Data Science*, 3(2):583–623.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77:49–61.

Kane, M., Price, N., Scotch, M., and Rabinowitz, P. (2014). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics*, 15(1).

Kim, J. K. and Rao, J. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96(4):917–932.

Klusowski, J. M. and Tian, P. M. (2024). Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537.

Krennmair, P. and Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(5):1865–1894.

Kuhn, M. (2022). *caret: Classification and Regression Training*. R package version 6.0-93.

Lohr, S., Hsu, V., and Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. In *Joint Statistical Meetings, Proceedings of the Survey Research Methods Section: American Statistical Association*, pages 2071–2085. American Statistical Association Alexandria, VA, USA.

McConville, K. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2):389–413.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881.

Michal, V., Wakefield, J., Schmidt, A. M., Cavanaugh, A., Robinson, B., and Baumgartner, J. (2023). Small area estimation with random forests and the lasso. *arXiv preprint arXiv:2308.15180*.

Nalenz, M., Rodemann, J., and Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine Learning*, pages 1–20.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*.

Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105.

Opsomer, J. and Miller, C. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Nonparametric Statistics*, 17(5):593–611.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In *Handbook of statistics*, volume 29, pages 455–487. Elsevier.

Phipps, P. and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics*, pages 772–794.

Qi, Y. (2012). *Random forests for bioinformatics*, pages 307–323. Springer.

Quinlan, J. (1993). Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning*, pages 236–243.

Robinson, P. M. and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā Ser. B*, 45(2):240–248.

Rogez, G., Rihan, J., Ramalingam, C., Orrite, C., and Torr, P. (2008). Randomized trees for human pose detection. In *Computer Vision and Pattern Recognition, CVPR, IEEE Conference on.*, pages 1–8.

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

Scornet, E. (2016). On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94(445):254–265.

Smucler, E., Rotnitzky, A., and Robins, J. M. (2019). A unifying approach for doubly-robust l1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*.

Stekhoven, D. J. and Buhlmann, P. (2011). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Therneau, T. and Atkinson, B. (2022). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.16.

Tipton, J., Opsomer, J., and Moisen, G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. *Remote sensing of environment*, 139:130–137.

Toth, D. and Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496):1626–1636.

Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77(i01).

Xu, T., Zhu, R., and Shao, X. (2024). On variance estimation of random forests with infinite-order u-statistics. *Electronic Journal of Statistics*, 18(1):2135–2207.

Yang, S. and Kim, J. K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. In *The Econometrics of Complex Survey Data: Theory and Applications*, pages 209–234. Emerald Publishing Limited.

Zhou, Z., Mentch, L., and Hooker, G. (2019). Asymptotic normality and variance estimation for supervised ensembles. *arXiv preprint arXiv:1912.01089*.