Model-assisted estimation through random forests in finite population sampling

Mehdi Dagdoug[(a)], Camelia Goga[(a)] and David Haziza[(b)]
(a) Université de Bourgogne Franche-Comté,
Laboratoire de Mathématiques de Besançon, Besançon, FRANCE
(b) University of Ottawa, Department of mathematics and statistics,
Ottawa, CANADA

Abstract

In surveys, the interest lies in estimating finite population parameters such as population totals and means. In most surveys, some auxiliary information is available at the estimation stage. This information may be incorporated in the estimation procedures to increase their precision. In this article, we use random forests to estimate the functional relationship between the survey variable and the auxiliary variables. In recent years, random forests have become attractive as National Statistical Offices have now access to a variety of data sources, potentially exhibiting a large number of observations on a large number of variables. We establish the theoretical properties of model-assisted procedures based on random forests and derive corresponding variance estimators. A model-calibration procedure for handling multiple survey variables is also discussed. The results of a simulation study suggest that the proposed point and estimation procedures perform well in term of bias, efficiency and coverage of normal-based confidence intervals, in a wide variety of settings. Finally, we apply the proposed methods using data on radio audiences collected by Médiamétrie, a French audience company.

Key words: Model-assisted approach; Model-calibration; Nonparametric regression; Random forest; Survey data; Variance estimation.

1 Introduction

Since the pioneering work of Särndal (1980), Robinson and Särndal (1983) and Särndal and Wright (1984), model-assisted estimation procedures have attracted a lot of attention in the literature; see also Särndal et al. (1992) for a comprehensive discussion

Mehdi Dagdoug's research was supported by grants of the region of Franche-Comté and Médiamétrie.

of the model-assisted approach. At the estimation stage, auxiliary information is often available and can be incorporated in the estimation procedures to increase the precision of the resulting point estimators. The model-assisted approach starts with postulating a working model, describing the relationship between a survey variable Y and a set of p auxiliary variables $X_1, X_2, \ldots, X_p$. The model is fitted to the sample observations to obtain predicted values, which then serve to build point estimators of population means/totals. Model-assisted estimators are asymptotically design-unbiased and design consistent, irrespective of whether or not the working model is correctly specified, which is an attractive feature; see Särndal et al. (1992) and Breidt and Opsomer (2017), among others. When the working model holds, model-assisted estimators are expected to be highly efficient. However, when the sample size is small, the use of model-assisted estimators requires some caution as they may suffer from small sample bias. In this article, we use random forests to estimate the functional relationship between Y and $X_1, X_2, \ldots, X_p$. In recent years, random forests have become attractive as National Statistical Offices have now access to a variety of data sources, potentially exhibiting a large number of observations on a large number of variables.

Consider a finite population $U = \{1, ..., k, ..., N\}$ of size N. We are interested in estimating the population total of a survey variable Y, $t_y = \sum_{k \in U} y_k$. We select a sample S, of size n, according to a sampling design $\mathcal{P}(S \mid \mathbf{Z}_U)$, where $\mathbf{Z}_U$ denotes the matrix of design information, available prior to sampling for all the population units. Let $\mathbf{I}_U = (I_1, ..., I_k, ..., I_N)^\top$ be the N-vector of sample selection indicators such that $I_k = 1$ if $k \in S$ and $I_k = 0$, otherwise. The first-order and second-order inclusion probabilities are given by $\pi_k = \mathbb{E}\left[I_k \mid \mathbf{Z}_U\right]$ and $\pi_{kl} = \mathbb{E}\left[I_k I_\ell \mid \mathbf{Z}_U\right]$, respectively.

A basic estimator of t_y is the well-known Horvitz-Thompson estimator given by

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

Provided that $\pi_k > 0$ for all $k \in U$, the estimator (1) is design-unbiased for t_y in the sense $\mathbb{E}\left[\widehat{t}_\pi \mid \mathbf{y}_U, \mathbf{Z}_U\right] = t_y$, where $\mathbf{y}_U = (y_1, y_2, ..., y_N)^\top$. The Horvitz-Thompson estimator makes no use of auxiliary information beyond what is already contained in the matrix $\mathbf{Z}_U$.

We assume that a vector $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kp})^\top$ of auxiliary variables is available for all $k \in U$. We also assume that y_k, $k \in U$, are independent realizations from a

working model ξ, often referred to as a superpopulation model:

$$\begin{aligned} \mathbb{E}\left[y_k \mid \mathbf{X}_k = \mathbf{x}_k\right] &= m(\mathbf{x}_k), \\ \mathbb{V}(y_k \mid \mathbf{X}_k = \mathbf{x}_k) &= \sigma^2 \nu(\mathbf{x}_k), \end{aligned} \tag{2}$$

where $m(\cdot)$ and $\nu(\cdot)$ are two unknown functions and σ^2 is an unknown parameter.

Suppose that Model (2) is fitted at the population level and let $\widetilde{m}(\mathbf{x}_k)$ be the population-level fit associated with unit k obtained by fitting a parametric or non-parametric procedure. This leads to the pseudo generalized difference estimator

$$\widehat{t}_{pgd} = \sum_{k \in U} \widetilde{m}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \widetilde{m}(\mathbf{x}_k)}{\pi_k}. \tag{3}$$

Because the values $\widetilde{m}(\mathbf{x}_k)$ do not involve the sample selection indicators $I_1, \ldots, I_N$, if follows that $\mathbb{E}\left[\widehat{t}_{pgd} \mid \mathbf{y}_U, \mathbf{Z}_U, \mathbf{X}_U\right] = t_y$, where $\mathbf{X}_U$ is the $N \times p$ matrix whose N rows are the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$. That is, the pseudo generalized difference estimator (3) is design-unbiased for t_y. In the sequel, we use the simpler notation $\mathbb{E}_p\left[\cdot\right]$ instead of $\mathbb{E}\left[\cdot|\mathbf{Z}_U, \mathbf{X}_U, \mathbf{y}_U\right]$ to denote the expectation operator with respect to the sampling design $\mathcal{P}(S|\mathbf{Z}_U)$. Similarly, the notation $\mathbb{V}_p\left[\cdot\right]$ is used to denote the design variance of an estimator.

Most often, the estimator (3) is unfeasible as the population-level fits $\widetilde{m}(\mathbf{x}_k)$ are unknown. Using the sample observations, we fit the working model and obtain the sample-level fits $\widehat{m}(\mathbf{x}_k)$. Replacing $\widetilde{m}(\mathbf{x}_k)$ with $\widehat{m}(\mathbf{x}_k)$ in (3), we obtain the so-called model-assisted estimator of t_y:

$$\widehat{t}_{ma} = \sum_{k \in U} \widehat{m}(\mathbf{x}_k) + \sum_{k \in S} \frac{y_k - \widehat{m}(\mathbf{x}_k)}{\pi_k}. \tag{4}$$

Unlike (3), the estimator (4) is no longer design-unbiased, but can be shown to be design-consistent for t_y for a relatively wide class of procedures $\widehat{m}(\cdot)$. The model-assisted estimator (4) is expressed as the sum of the population total of the predictions $\widehat{m}(\mathbf{x}_k)$ and an adjustment term that can be viewed as a protection against model-misspecification.

If $\widehat{m}(\mathbf{x}_k) = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}$ with coefficients estimated by weighted least squares, the estimator (4) reduces to the well-known generalized regression (GREG) estimator; e.g., see Särndal et al. (1992, Chap. 6). Model-assisted estimators based on generalized

linear models were considered by Lehtonen and Veijanen (1998) and Firth and Bennett (1998), among others. There are some practical issues associated with the use of a parametric model such as linear and generalized linear models: they may lead to inefficient estimators if the function $m(\cdot)$ is misspecified or if the model fails to include interactions or predictors that account for curvature (e.g., quadratic and cubic terms). In contrast, nonparametric procedures are robust to model misspeficiation, which is a desirable property. A number of nonparametric model-assisted estimation procedures have been studied in the last two decades: local polynomial regression (Breidt and Opsomer, 2000), B-splines (Goga, 2005) and penalized B-splines (Goga and Ruiz-Gazen, 2014), penalized splines (Breidt et al., 2005; McConville and Breidt, 2013), neural nets (Montanari and Ranalli, 2005), generalized additive models (Opsomer et al., 2007), nonparametric additive models (Wang and Wang, 2011) and regression trees (Toth and Eltinge, 2011; McConville and Toth, 2019).

In this paper, we propose a new class of model-assisted estimators of t_y based on random forests (RF). Generally speaking, RF is an ensemble method that trains a (large) number of trees and combines them to produce more accurate predictions than a single regression tree would. Trees define a class of algorithms that recursively split the p-dimensional predictor space into distinct and non-overlapping regions. In other words, a tree algorithm generates a partition of regions or hyperrectangles of $\mathbb{R}^p$. For an observation belonging to a given region, the prediction is simply obtained by averaging the y-values associated with the units belonging to the same region. While regression trees are easy to interpret and allow the user to visualize the partition (Hastie et al., 2011, pp. 306), they may suffer from a high model variance, hence their qualification of "weak learners". A number of tree-based procedures have been proposed with the aim of improving the predictive performances of regression trees, including pruning (Breiman et al., 1984), Bayesian regression trees (Chipman et al., 1998), gradient boosting (Friedman, 2001) and RF (Breiman, 2001).

Several empirical studies suggest that RF can outperform state-of-the-art prediction models; see e.g. Han et al. (2018), Hamza and Larocque (2005), Díaz-Uriarte and de Andrés (2006). RF are widely used due to their predictive performances and their ability to handle small sample sizes with a large number of predictors (Scornet, 2016b). Also, RF algorithms can be parallelized, leading to a decrease in the training time. RF

have been applied in a wide variety of fields, 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). In recent years, neural networks and deep learning algorithms have attracted a lot of attention and have been shown to be effective in a wide range of applications involving mostly unstructured data, such as speech recognition, image reconstruction and text translation; see Najafabadi et al. (2015) and the references therein for a review on the topic. However, to exhibit high levels of performance, deep learning algorithms typically require huge amounts of data (Najafabadi et al., 2015; Arnould et al., 2020). This is seldom the case in surveys as most data sets consist of structured data consisting of (at most) a few hundred thousand observations and a few hundred survey variables. For an empirical comparison of RF and neural networks, see Han et al. (2018). Finally, unlike RF algorithms that require the specification of a small number of hyper-parameters (see Section 6.3), gradient boosting, Bayesian regression trees or deep learning approaches depend upon the complex choice of a large number of hyper-parameters (Bergstra et al., 2011).

To the best of our knowledge, only little is known about the theoretical properties of RF based on the original algorithm of Breiman (2001). Often, the theoretical investigations are made at the expense of simplifying assumptions; see for instance Biau et al. (2008) and Biau (2012). Two notable exceptions are Wager (2014) and Scornet et al. (2015) who established the theoretical properties of an algorithm closely related to that of Breiman (2001). In a finite population setting, the theoretical properties of RF algorithms have yet to be established, even in the ideal situation of 100% response. This paper aims to fill this important gap. While we are mostly concerned with RF for regression, we can easily extend our methods to the case of RF for classification. Some recent empirical studies on the performance of RF for complex survey data can be found in Tipton et al. (2013), Buskirk and Kolenikov (2015), De Moliner and Goga (2018) and Kern et al. (2019).

The rest of the paper is organized as follows. Regression trees and RF are presented in Section 2. In Section 3, we suggest two classes of model-assisted estimators based on random forests: the first is based on partitions built at the population level, while the second class is based on partitions built at the sample level. In Section 4, we

establish the theoretical properties of model-assisted estimators based on RF and derive corresponding variance estimators. In Section 5, we describe a model-calibration procedure for handling multiple survey variables. In Sections 6.1-6.3, the finite sample properties of the proposed point and variance estimation procedures are evaluated through a simulation study, and in Section 6.4, we apply the proposed methods using data on radio audiences collected by Médiamétrie, a French audience company. The paper ends with some final remarks in Section 7. Proofs of major results and further technical details are relegated to the Appendix and the Supplementary Material.

2 Regression trees and random forests

2.1 Regression trees

The original RF uses regression trees based on the classification and regression tree algorithm (CART) of Breiman et al. (1984), whereby the partition of the predictor space is generated by a greedy recursive algorithm. In this paper, we focus on the CART algorithm for regression, designed for handling quantitative survey variables Y, but our methods also applies to the case of binary survey variables. With regression trees, these estimated probabilities always lie between 0 and 1, which is a desirable feature. Alternative criteria may be used with binary variables, such as the Gini impurity or the entropy instead of the CART regression criterion (Hastie et al., 2011, Chapter 9). The CART algorithm for regression searches for the splitting variable and the splitting position (i.e., the coordinates on the predictor space where to split) for which the difference in empirical variance in the node before and after splitting is maximized. As a starting point, we consider the hypothetical situation, where y_k and $\mathbf{x}_k$ are observed for all $k \in U$ and assume that the regression tree is fitted at the population level. We use the generic notation A to denote a node with cardinality $\#(A)$ considered for the next split, and $\mathcal{C}_A$ to denote the set of possible splits in the node A, which corresponds to the set of all possible pairs $(j, z) = (\text{variable}, \text{position})$. This splitting process is performed by searching for the best split (j^*, z^*) for which the following empirical CART population criterion is maximized:

$$L_N(j,z) = \frac{1}{\#(A)} \sum_{k \in U} \mathbb{1}_{\mathbf{x}_k \in A} \left\{ (y_k - \bar{y}_A)^2 - \left(y_k - \bar{y}_{A_L} \mathbb{1}_{x_{kj} < z} - \bar{y}_{A_R} \mathbb{1}_{x_{kj} \geq z} \right)^2 \right\}, \qquad (5)$$

where $A_L = \{k \in A; x_{kj} < z\}$, $A_R = \{k \in A; x_{kj} \geqslant z\}$ and $\bar{y}_A$ is the average of the y-values of units belonging to A. The best cut is always performed in the middle of two consecutive data points. In practice, it is common to impose a minimal number of observations N_0 (say) in each terminal node. In this case, the splitting process is performed until an additional split generates a terminal node with fewer observations than N_0.

The splitting process leads to the set

$$\mathcal{P}_U = \left\{ A_1^{(U)}, \ldots, A_j^{(U)}, \ldots, A_{J_U}^{(U)} \right\} \tag{6}$$

of J_U hyperrectangles of $\mathbb{R}^p$ such that $A_j^{(U)} \bigcap A_{j'}^{(U)} = \emptyset$, for all $j \neq j' \in \{1, 2, \ldots, J_U\}$ and $\bigcup_{j=1}^{J_U} A_j^{(U)} = \mathbb{R}^p$. Thus, the set $\mathcal{P}_U$ defines a partition of $\mathbb{R}^p$, whose elements are called the terminal nodes. We use the generic notation $A^{(U)}(\mathbf{x}_k)$ to denote a terminal node belonging to the partition $\mathcal{P}_U$ given in (6) and that contains $\mathbf{x}_k$.

Figure 1 below illustrates how the recursive splitting procedure creates a partition in the simple case of two auxiliary variables X_1 and X_2, based on 5 splits. Each grey rotated square represents a split (variable, position) performed at some position along one of the two auxiliary variables, X_1 or X_2. The white ellipses represent the 6 terminal nodes, also represented by the scatter plot on the right; see also Biau and Devroye (2014) for a similar illustration.

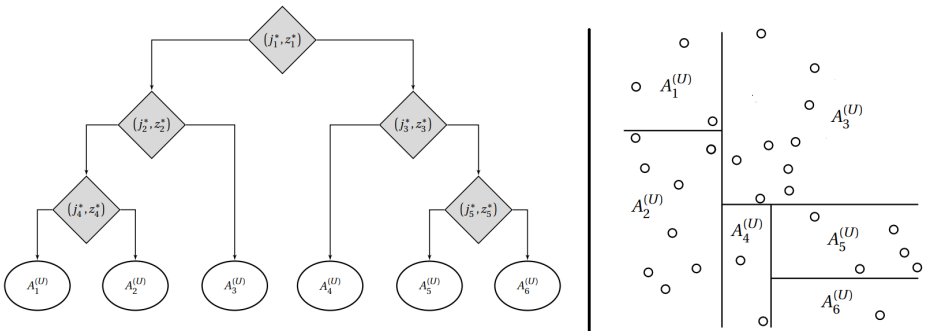

Figure 1: A regression tree (left) and the corresponding partition of $\mathbb{R}^2$ (right).

The prediction $\widetilde{m}_{tree}(\mathbf{x}_k)$ at the point $\mathbf{x}_k$ is simply defined as the average of the y-values of population individuals ℓ such that $\mathbf{x}_\ell$ belongs to $A^{(U)}(\mathbf{x}_k)$:

$$\widetilde{m}_{tree}(\mathbf{x}_k) = \sum_{\ell \in U} \frac{\mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}(\mathbf{x}_k)} y_\ell}{\widetilde{N}(\mathbf{x}_k)}, \tag{7}$$

where $\widetilde{N}(\mathbf{x}_k) = \sum_{\ell \in U} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}(\mathbf{x}_k)}$ denotes,the number of units belonging to the terminal node $A^{(U)}(\mathbf{x}_k)$. Given the partition $\mathcal{P}_U$, the population-level fit $\widetilde{m}_{tree}(\mathbf{x}_k)$ may be viewed as the least squares type prediction obtained by fitting a one-way ANOVA model with Y as the response variable and the node membership indicators $\{\mathbb{1}_{\mathbf{x}_k \in A_j^{(U)}}\}_{j=1}^{J_U}$ as the set of explanatory variables; see (Hastie et al., 2011, Chapter 9) and the Supplementary Material for more details.

2.2 Random forests

To introduce random forests (RF) in a finite population setting, we again assume that y_k and $\mathbf{x}_k$ are observed for all $k \in U$. RF are based on a (large) number B (say) of regression trees. The prediction attached to unit k is defined as the average of the predictions produced by each of the B regression trees. That is,

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \widetilde{m}_{tree}^{(b)}(\mathbf{x}_k),$$

where $\widetilde{m}_{tree}^{(b)}(\mathbf{x}_k)$ is the predicted value attached to unit k obtained from the bth regression tree, $b = 1, \ldots, B$.

Obviously, if $\widetilde{m}_{tree}^{(1)}(\mathbf{x}_k) = \ldots = \widetilde{m}_{tree}^{(B)}(\mathbf{x}_k)$, then $\widetilde{m}_{rf}(\mathbf{x}_k) = \widetilde{m}_{tree}^{(1)}(\mathbf{x}_k)$. Such a situation would occur if each regression tree uses a deterministic splitting criterion in (5), which would lead to B identical partitions of $\mathbb{R}^p$. To cope with this issue, some amount of randomization is introduced in the tree building process, leading to B different predictions of $m(\cdot)$. The original algorithm of Breiman (2001) is implemented as follows:

1. Select B bootstrap data sets with replacement from the population data set, $D_U = \{(\mathbf{x}_k, y_k)\}_{k \in U}$, each data set containing N pairs of the form $(\mathbf{x}_k, y_k)$;

2. Fit a regression tree on each bootstrap data set. Before each split is performed, m_{try} predictors are selected randomly and without replacement from the full set of p predictors. The m_{try} selected predictors are the split candidates to be considered for searching the best split in (5).

The algorithm stops when each terminal node contains less than a predetermined number of observations. This procedure leads to a set $\widetilde{\mathcal{P}}_U = \left\{\mathcal{P}_U^{(1)}, \mathcal{P}_U^{(2)}, \ldots, \mathcal{P}_U^{(B)}\right\}$

of B different partitions of $\mathbb{R}^p$, each of the form (6). The randomization used in the tree building process is denoted by the random variable $\theta^{(U)}$, assumed to belong to some measurable space $(\Theta, \mathcal{F})$ and independent of the data (Biau and Scornet, 2016). Let $\theta_b^{(U)}$ be the random variable associated with the bth tree. The random variables $\theta_b^{(U)}, b = 1, \ldots, B$, are assumed to be independent and their distribution is identical to that of the generic random variable $\theta^{(U)}$. In the RF algorithm of Breiman, the randomization is induced by the selection (with replacement) of observations in Step 1 of the above algorithm and the random selection of split variables in Step 2 of the above algorithm. A number of RF algorithms algorithms have been considered in the literature. For example, (Biau et al., 2008; Scornet, 2016a) considered a simple RF algorithm called the uniform random forest (URF) algorithm. In the URF algorithm, a variable is selected with equal probability among the initial p predictors at each node and a split position is chosen uniformly in the node along the direction of the selected variable. The algorithm stops when each terminal node has a predetermined number of cuts. In this case, the randomization $\theta_b^{(U)}$ is characterized by the random selections of the node, the split variable and the location. For more details on RF algorithms, the reader is referred to Geurts et al. (2006), Biau et al. (2008), Biau (2012), Genuer (2012), Scornet (2016a), among others. In the sequel, unless stated otherwise, we assume that the observations in Step 1 of the above algorithm are selected without replacement (Scornet, 2017), which we will refer to as subsampling. Also, for more generality, the splitting criterion is left unspecified.

Let $\widetilde{m}_{tree}^{(1)}(\cdot, \theta_1^{(U)}), \ldots, \widetilde{m}_{tree}^{(B)}(\cdot, \theta_B^{(U)})$, denote the predictions obtained with the B stochastic or randomized regression trees. The RF prediction attached to unit k is defined as a bagged estimator of B trees:

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \widetilde{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(U)}). \tag{8}$$

It is worth pointing out that considering a new set of predictors at each split leads to B trees which are less correlated with each other; that is, trees that are quite different from one another. As a result, the RF may lead to substantial gains in precision compared to a single tree (James et al., 2015, Chapter 8). The number of predictors selected at each split, denoted by m_{try}, is thus an important tuning parameter in the RF algorithm. In practice, the choice $m_{try} = \sqrt{p}$ seems to give good results, in general.

In Section 6.3, we assess the impact of m_{try} through a simulation study.

For any RF algorithm, the prediction at the point $\mathbf{x}_k$ in (8) can also be expressed as

$$\widetilde{m}_{rf}(\mathbf{x}_k) = \sum_{\ell \in U} \widetilde{W}_\ell(\mathbf{x}_k) y_\ell, \tag{9}$$

where

$$\widetilde{W}_\ell(\mathbf{x}_k) = \frac{1}{B} \sum_{b=1}^{B} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}\left(\mathbf{x}_k, \theta_b^{(U)}\right)}}{\widetilde{N}(\mathbf{x}_k, \theta_b^{(U)})} \tag{10}$$

is a prediction weight attached to unit k with $\widetilde{N}(\mathbf{x}_k, \theta_b^{(U)}) = \sum_{\ell \in U} \psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}(\mathbf{x}_k, \theta_b^{(U)})}$ denoting the number of observations belonging to the terminal node $A^{(U)}$ containing $\mathbf{x}_k$ in the bth regression tree. The random variables $\psi_\ell^{(b,U)}$ in (10) depend on the resampling mechanism used in the RF algorithm and depend on $\theta_b^{(U)}$, but are independent of the sampling design $\mathcal{P}(S \mid \mathbf{Z}_U)$. In the case of subsampling, the random variables $\psi_\ell^{(b,U)}$ follow a Bernoulli distribution, $\psi_\ell^{(b,U)} \sim \mathcal{B}(N'/N)$, where N' denotes the number of units in each subsample. Note that the prediction $\widetilde{m}_{rf}$ in (9) can be computed for either a continuous or a categorical y-variable. In the latter case, the prediction $\widetilde{m}_{rf}$ in (9) corresponds to the population proportion of units who belong to a given category computed over the B trees.

Proposition 2.1. *Consider the predictor weights $\widetilde{W}_\ell(\mathbf{x}_k)$ given in (10).*

i) The weights $\widetilde{W}_\ell(\mathbf{x}_k)$ are uniformly bounded. That is,

$$0 < \widetilde{W}_\ell(\mathbf{x}_k) \leqslant c N_0^{-1}$$

for all $\ell \in U$ and all $\mathbf{x}_k \in \mathbb{R}^p$, where c is a positive constant that does not depend either on k,ℓ, or N_0, the minimal number of observations in the terminal nodes.

ii) The weight functions sum up to one; that is, $\sum_{\ell \in U} \widetilde{W}_\ell(\mathbf{x}_k) = 1$ for all $\mathbf{x}_k \in \mathbb{R}^p$.

The proof of Proposition 2.1 is given in the Appendix.

3 Model-assisted estimation: Random forests

In Section 2, we assumed that y_k and $\mathbf{x}_k$ were observed for all $k \in U$, which led to the population-level fits $\widetilde{m}_{tree}(\mathbf{x}_k)$ and $\widetilde{m}_{rf}(\mathbf{x}_k)$ given by (7) and (9), respectively. However, both (7) and (9) cannot be computed in practice as the y-values are observed

only for $k \in S$. Moreover, the regression trees in Sections 2.1 and 2.2 were based on partitions built recursively at the population level so as to optimize the population criterion (5). As a result, these partitions depend on the vector of predictors $\{\mathbf{x}_k\}_{k\in U}$ but also on the unknown population values $\{y_k\}_{k\in U}$. While the former type of dependency is inherent to most parametric and nonparametric procedures, the latter is absent in many commonly used parametric and nonparametric procedures such as spline procedures (Breidt and Opsomer, 2000; Goga, 2005; Goga and Ruiz-Gazen, 2014; Breidt et al., 2005; McConville and Breidt, 2013). Due to the dependency on the unknown population values $\{y_k\}_{k\in U}$, establishing the theoretical properties of model-assisted estimators based on RF is more challenging.

For these reasons, in Section 3.1, we start by considering the simpler case of population partitions obtained using a variable Y^*, assumed to be closely related to Y and available for all $k \in U$. While this assumption is somehow strong and not tenable in many practical situations, it provides some insights on how to tackle the problem in the presence of Y-dependency. Algorithms allowing to get rid of the Y-dependency have been suggested in the random-forest literature; see e.g. Biau et al. (2008), Biau (2012) or Devroye et al. (1996, Chap. 20). Sample-based partitions are considered in Section 3.2.

3.1 Model-assisted estimation: Population-based partitions

In this section, we consider the case of a splitting criterion that does not depend on the data $\{y_k\}_{k\in s}$. We consider a variable Y^* assumed to be closely related to Y and such that the values y_k^* are available for all $k \in U$. We seek population partitions $\widetilde{\mathcal{P}}_U^*$, independent of the survey variable Y, that maximize the following criterion:

$$L_N^*(j,z) = \frac{1}{\#(A)} \sum_{k\in U} \mathbb{1}_{\mathbf{x}_k\in A} \left\{ (y_k^* - \overline{y}_A^*)^2 - \left(y_k^* - \overline{y}_{A_L}^* \mathbb{1}_{x_{kj}<z} - \overline{y}_{A_R}^* \mathbb{1}_{x_{kj}\geq z}\right)^2 \right\}, \quad (11)$$

where A_R, A_L are as in (5) and $\overline{y^*}_A$ is the average of the y^*-values for the units belonging to a node A.

Based on (11), the population-level fit at the point $\mathbf{x}_k$ is given by

$$\widetilde{m}_{rf}^*(\mathbf{x}_k) = \sum_{\ell\in U} \widetilde{W}_\ell^*(\mathbf{x}_k) y_\ell, \quad (12)$$

where the weights $\widetilde{W}_\ell^*(\mathbf{x}_k)$ in (12) are obtained from (10) by replacing $A^{(U)}$ with $A^{*(U)}$, a generic member of the partition $\widetilde{\mathcal{P}}_U^*$.

The weights $\{\widetilde{W}_\ell^*(\cdot)\}_{\ell\in U}$ in (12) are known for all $\ell \in U$ and are independent of Y. Since $\widetilde{m}_{rf}^*(\mathbf{x}_k)$ in (12) requires the y-values for all the population units, it cannot be computed. A simple solution consists of replacing the population total on the right hand-side of (12) by its corresponding Horvitz–Thompson estimator, which leads to

$$\widehat{m}_{rf}^*(\mathbf{x}_k) = \sum_{\ell\in S} \frac{\widetilde{W}_\ell^*(\mathbf{x}_k) y_\ell}{\pi_\ell}. \tag{13}$$

A model-assisted estimator of t_y based on population RF is obtained by plugging $\widehat{m}_{rf}^*(\mathbf{x}_k)$ in (3):

$$\widehat{t}_{rf}^* = \sum_{k\in U} \widehat{m}_{rf}^*(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widehat{m}_{rf}^*(\mathbf{x}_k)}{\pi_k}. \tag{14}$$

Proposition 3.1. *The RF estimator given in (14) can be expressed as*

$$\widehat{t}_{rf}^* = \sum_{k\in S} w_{ks} y_k,$$

where the weights w_{ks} are given by

$$w_{ks} = \frac{1}{\pi_k}\left\{1 + \sum_{\ell\in U} \widetilde{W}_k^*(\mathbf{x}_\ell)\left(1 - \frac{I_\ell}{\pi_\ell}\right)\right\}. \quad k \in S \tag{15}$$

Proof. By rearranging the sums, we get:

$$\begin{aligned}\widehat{t}_{rf}^* &= \sum_{k\in S}\frac{y_k}{\pi_k} + \sum_{\ell\in U}\left(1 - \frac{I_\ell}{\pi_\ell}\right)\widehat{m}_{rf}^*(\mathbf{x}_\ell) = \sum_{k\in S}\frac{y_k}{\pi_k} + \sum_{\ell\in U}\left(1 - \frac{I_\ell}{\pi_\ell}\right)\left(\sum_{k\in S}\widetilde{W}_k^*(\mathbf{x}_\ell)\frac{y_k}{\pi_k}\right) \\ &= \sum_{k\in S}\left\{1 + \sum_{\ell\in U}\left(1 - \frac{I_\ell}{\pi_\ell}\right)\widetilde{W}_k^*(\mathbf{x}_\ell)\right\}\frac{y_k}{\pi_k}.\end{aligned}$$

■

Since the partitions $\widetilde{\mathcal{P}}_U^*$, are independent of both the survey variable Y and the sample S, the weights w_{ks} given by (15) depend on the sample only through the sample selection indicators $I_\ell, \ell \in U$, but are independent of Y. As a result, these weights may be used to estimate the population total of any survey variable, which is an attractive feature in multipurpose surveys. However, for RF algorithms based on the splitting criterion in (11), we expect the weights w_{ks} to be efficient whenever the survey variable

Y is highly correlated to the variable Y^*. In multipurpose surveys where the survey variables are not necessarily correlated with one another, it may be preferable to use a splitting criterion that depends on the data $\{\mathbf{x}_k\}_{k\in U}$ as done in quantile random forests (Devroye et al., 1996; Scornet, 2016a).

3.2 Model-assisted estimation: Sample-based partitions

In this section, we seek sample partitions $\widehat{\mathcal{P}}_S = \left\{\widehat{\mathcal{P}}_S^{(1)}, \ldots, \widehat{\mathcal{P}}_S^{(b)}, \ldots, \widehat{\mathcal{P}}_S^{(B)}\right\}$ using

$$L_n(j,z) \;=\; \frac{1}{\#(A)}\sum_{k\in S}\mathbb{1}_{\mathbf{x}_k\in A}\left\{(y_k-\bar{y}_A)^2-\left(y_k-\bar{y}_{A_L}\mathbb{1}_{x_{kj}<z}-\bar{y}_{A_R}\mathbb{1}_{x_{kj}\geq z}\right)^2\right\}. \quad (16)$$

Based on the partition $\widehat{\mathcal{P}}_S$, we obtain the sample-level fits

$$\widehat{m}_{rf}(\mathbf{x}_k) = \sum_{\ell\in S}\frac{\widehat{W}_\ell(\mathbf{x}_k)y_\ell}{\pi_\ell}, \quad (17)$$

where

$$\widehat{W}_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B}\frac{\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A^{(S)}\left(\mathbf{x}_k,\theta_b^{(S)}\right)}}{\widehat{N}(\mathbf{x}_k,\theta_b^{(S)})}, \quad \ell\in S, \quad (18)$$

and $\widehat{N}(\mathbf{x}_k,\theta_b^{(S)}) = \sum_{\ell\in U} I_\ell\pi_\ell^{-1}\psi_\ell^{(b,S)}\mathbb{1}_{\mathbf{x}_\ell\in A^{(S)}\left(\mathbf{x}_k,\theta_b^{(S)}\right)}$ denotes the estimated number of observations in the terminal node $A^{(S)}$ containing $\mathbf{x}_k$ in the bth regression tree. The variable $\psi_\ell^{(b,S)}$ indicates whether or not unit ℓ has been selected in the bth sub-sample and is such that $\psi_\ell^{(b,S)}\sim\mathcal{B}(n'/n)$ for RF based on subsampling, where n' denotes the number of units in each sub-sample.

Plugging $\widehat{m}_{rf}(\cdot)$ in (4) leads to the RF model-assisted estimator

$$\widehat{t}_{rf} = \sum_{k\in U}\widehat{m}_{rf}(\mathbf{x}_k) + \sum_{k\in S}\frac{y_k-\widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k}. \quad (19)$$

Using similar arguments to those used in the proof of Proposition 3.1, we can show that $\widehat{t}_{rf}$ can be expressed as

$$\widehat{t}_{rf} = \sum_{k\in S} w'_{ks}y_k,$$

where the weights w'_{ks} are given by

$$w'_{ks} = \frac{1}{\pi_k}\left\{1+\sum_{\ell\in U}\widehat{W}_k(\mathbf{x}_\ell)\left(1-\frac{I_\ell}{\pi_\ell}\right)\right\}, \quad k\in S. \quad (20)$$

Noting that $\sum_{k\in S}\widehat{W}_k(\mathbf{x}_\ell)\pi_k^{-1} = 1$ for all $\ell\in U$, it follows from (20) that $\sum_{k\in S} w'_{ks} = N$ for every sample S. That is, the sum of the weights w'_{ks} match

the population size N perfectly, a desirable property shared by other nonparametric model-assisted estimators (Goga, 2005; Goga and Ruiz-Gazen, 2014; Breidt et al., 2005). Unlike the weights w_{ks} in (15), the weights w'_{ks} depend on both the sample selection indicators $I_\ell, \ell \in U$, and the partition $\widehat{\mathcal{P}}_S$ that varies from one sample to another. This is due to the fact that the nodes $A^{(S)}$ are constructed so as to optimize the sample criterion (16). For this reason, the weights $w'_{ks}, k \in S$, are variable specific in the sense that depend on the survey variable Y. To cope with this issue, we describe a model calibration procedure in Section 5 for handling multiple survey variables while producing a single set of weights.

Remark 3.1. *In practice, the variables $\psi_k^{(b,S)}$ in (18) are not generated for the units outside the sample. However, at least conceptually, nothing precludes defining these variables for $k \in U \setminus S$. For $k \in U \setminus S$, we set $\psi_k^{(b,S)} \sim \mathcal{B}\left((N'-n')/(N-n)\right)$ so that $\sum_{k\in U} \psi_k^{(b,S)} = N'$. Defining the variables $\psi_k^{(b,S)}$ for units outside the sample will have no effect on the predictions $\widehat{m}_{rf}(\cdot)$ associated with the sample units since $I_k = 0$ for $k \in U \setminus S$. This construction will prove useful in establishing the asymptotic properties of the proposed procedures; see Section 4.*

As for the RF prediction built at the population level described in Section 2.2, the prediction $\widehat{m}_{rf}(\mathbf{x}_k)$ in (17) can be expressed as a bagged predictor (Hastie et al., 2011). That is,

$$\widehat{m}_{rf}(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}),$$

where $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}) = \sum_{\ell\in S} \pi_\ell^{-1}\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(S)}\left(\mathbf{x}_k, \theta_b^{(S)}\right)} y_\ell / \widehat{N}(\mathbf{x}_k, \theta_b^{(S)})$ is the prediction associated with unit k based on the bth stochastic regression tree. The model-assisted estimator $\widehat{t}_{rf}$ given by (19) can thus be viewed as a bagged estimator:

$$\widehat{t}_{rf} = \frac{1}{B}\sum_{b=1}^{B} \widehat{t}_{tree}^{(b)}(\theta_b^{(S)}),$$

where

$$\widehat{t}_{tree}^{(b)}(\theta_b^{(S)}) = \sum_{k\in U} \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}) + \sum_{k\in S} \frac{y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})}{\pi_k}$$

is the model-assisted estimator of t_y based on the bth stochastic regression tree. As in the case of regression trees built at the population level (see Section 2.1), given the partition $\widehat{\mathcal{P}}_S^{(b)} = \{A_j^{(bS)}\}_{j=1}^{J_{bS}}$, the predictions $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})$ are least squares type

predictions obtained by fitting the one-way ANOVA model with Y as the response and the node membership indicators $\{\mathbb{1}_{\mathbf{x}_k \in A_j^{(bS)}}\}_{j=1}^{J_{bS}}$ as the explanatory variables; see the proof of Proposition 3.3 and the Supplementary Material for more details. As a result, the estimator $\widehat{t}_{tree}^{(b)}(\theta_b^{(S)})$ is related to the customary post-stratified estimator (Särndal et al., 1992).

Under mild assumptions, Proposition 3.2 below shows that bagging improves the efficiency of model-assisted estimators. This is similar to what is encountered in the classical RF literature (Hastie et al., 2011).

Proposition 3.2. *Let $\widehat{t}^{(1)}, \ldots, \widehat{t}^{(b)}, \ldots, \widehat{t}^{(B)}$ be a sequence of model-assisted estimators of t_y and let $\widehat{t} = B^{-1}\sum_{b=1}^{B}\widehat{t}^{(b)}$ be a bagged estimator. Assuming that the $\widehat{t}^{(b)}$'s have approximately the same design bias and design variance, then, for B large enough:*

$$MSE_p(\widehat{t}) - MSE_p(\widehat{t}^{(1)}) \leqslant \mathbb{V}_p(\widehat{t}^{(1)})\left(\max_{b\neq b'}\left|\mathbb{C}or_p\left(\widehat{t}^{(b)}, \widehat{t}^{(b')}\right)\right| - 1\right) \leqslant 0,$$

where $MSE_p(\cdot)$ and $\mathbb{C}or_p(\cdot)$ denote the mean squared error and correlation operators with respect to the sampling design.

The proof of Proposition 3.2 is given in the Appendix. We end this section by giving an alternative expression for $\widehat{t}_{rf}$.

Proposition 3.3. *The RF estimator $\widehat{t}_{rf}$ given by (19) can be written as*

$$\widehat{t}_{rf} = \sum_{k\in U}\widehat{m}_{rf}(\mathbf{x}_k) + \frac{1}{B}\sum_{b=1}^{B}\sum_{k\in S}\frac{\left(1-\psi_k^{(b,S)}\right)\left(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})\right)}{\pi_k}, \tag{21}$$

where $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})$ is the predictor associated with unit k based on the bth stochastic regression tree.

The proof of Proposition 3.3 is given in the Appendix. It follows from Proposition 3.3, that the second term on the right hand-side of (21) vanishes if $\psi_k^{(b,S)} = 1$ for all $k \in S$. That is, the estimator $\widehat{t}_{rf}$ reduces to the so-called projection form (Särndal et al., 1992; Breidt et al., 2005; Goga, 2005)

$$\widehat{t}_{rf} = \sum_{k\in U}\widehat{m}_{rf}(\mathbf{x}_k)$$

if the RF algorithm does not involve a resampling mechanism. In addition, the second term on the right hand-side of (21) vanishes if $y_k = c$ for all k, for some $c \in \mathbb{R}$ or

if the trees in the forest are fully grown (i.e., each terminal node contains a single observation), which implies that the observations y_k and the corresponding prediction $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})$ coincide. When the estimator $\widehat{t}_{rf}$ can be expressed in the projection form, the weights w'_{ks} given by (20) are always positive and cannot exceed the number of terminal nodes from the largest tree of the forest.

In practice, a resampling mechanism is typically used with RF algorithms. In this case, the second term on the right hand-side of (21) does not vanish and is equal to the weighted sum of residuals computed for the non-resampled units, also called the *out-of-bag* individuals (James et al., 2015, Chapter 8), from each of the B trees. The second term on the right hand-side of (21) can then be viewed as a correction term which brings additional information from the units not used in computing the predictions $\widehat{m}_{tree}^{(b)}(\cdot, \theta_b^{(S)}), b = 1, \ldots, B$.

4 Asymptotic properties

To establish the asymptotic properties of the proposed estimators and to derive the associated variance estimators, we consider the asymptotic framework of Isaki and Fuller (1982). We start with an increasing sequence of embedded finite populations $\{U_v\}_{v\in\mathbb{N}}$ of size $\{N_v\}_{v\in\mathbb{N}}$. In each finite population U_v, a sample of size n_v is selected according to a sampling design $\mathcal{P}_v(S_v = s_v \mid \mathbf{Z}_U)$. While the finite populations are assumed to be embedded, we do not require this property to hold for the samples $\{S_v\}_{v\in\mathbb{N}}$. This asymptotic framework assumes that v goes to infinity, so that both the finite population sizes and the samples sizes go to infinity. To improve readability, we shall use the subscript v only in the quantities U_v, N_v and n_v; quantities such as $\pi_{k,v}$ shall be denoted simply as π_k.

Assumptions: RF model-assisted estimator $\widehat{t}_{rf}^*$

We make the following assumptions:

(H1) There exists a positive constant C such that $\sup_{k\in U_v} |y_k| \leqslant C < \infty$.

(H2) We assume that $\lim_{v\to\infty} \frac{n_v}{N_v} = \pi \in (0, 1)$.

(H3) There exist positive constants λ and λ^* such that $\min_{k\in U_v} \pi_k \geqslant \lambda > 0$ and $\min_{k,\ell\in U_v} \pi_{k\ell} \geqslant \lambda^* > 0$. Also, we assume that $\limsup_{v\to\infty} n_v \max_{k\neq\ell\in U_v} |\pi_{k\ell} - \pi_k\pi_\ell| < \infty$.

Assumptions (H1)-(H3) have been extensively used in parametric, nonparametric and functional model-assisted estimation (Robinson and Särndal, 1983; Breidt and Opsomer, 2000; Breidt et al., 2005; Goga, 2005; Goga and Ruiz-Gazen, 2014; Cardot et al., 2013). Assumption (H1) implies that the survey variable Y is uniformly bounded (Breidt and Opsomer, 2000; Cardot et al., 2010). Assumptions (H2) and (H3) deal with the first and second order inclusion probabilities and they are satisfied for the classical fixed-size sampling designs; see for example, Robinson and Särndal (1983) and Breidt and Opsomer (2000). Furthermore, we assume that the minimum number of observations N_{0v} in a terminal nodes is growing to infinity and we make the following additional assumption

(C1) The number of subsampled elements N'_v is such that $\lim_{v\to\infty} N'_v/N_v \in (0;1]$.

This assumption requires that the number N'_v of elements in each subsample increases at the same speed as the population size N_v, allowing each terminal node to have at least N_{0v} observations.

Assumptions: RF model-assisted estimator $\widehat{t}_{rf}$

In addition to the above assumptions, we make the following assumptions to establish the asymptotic properties of $\hat{t}_{rf}$ given by (19).

(H4) There exists a positive constant C_1 such that $n_v \max_{k\neq\ell\in U_v} \left|\mathbb{E}_p\left\{(I_k - \pi_k)(I_\ell - \pi_\ell)|\widehat{\mathcal{P}}_S\right\}\right| \leqslant C_1$.

(H5) The random forests based on population partitions and those based on sample partitions are such that, for all $\mathbf{x} \in \mathbb{R}^p$:

$$\mathbb{E}_p\left(\widehat{\widetilde{m}}_{rf}(\mathbf{x}) - \widetilde{m}_{rf}(\mathbf{x})\right)^2 = o(1),$$

where $\widehat{\widetilde{m}}_{rf}(\mathbf{x}) = \sum_{\ell\in U_v} \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell\in A^{(S)}\left(\mathbf{x},\theta_b^{(S)}\right)} y_\ell}{\widehat{\widetilde{N}}(\mathbf{x},\theta_b^{(S)})}$ with $\widehat{\widetilde{N}}(\mathbf{x},\theta_b^{(S)}) = \sum_{\ell\in U_v} \psi_\ell^{(b,S)} \mathbb{1}_{\mathbf{x}_\ell\in A^{(S)}\left(\mathbf{x},\theta_b^{(S)}\right)}$.

Assumption (H4) is similar to that used by Toth and Eltinge (2011) and McConville

and Toth (2019); it requires that, as the sample and population size grow, the influence of extreme observations on the sample partitions decreases. Assumption (H5) requires that the average number of elements at the population level in the sample partitions converges to the average number of population elements in the population partitions. It implicitly assumes that the sample partitions converge to the population partitions. A similar result was established in Toth and Eltinge (2011) in the case of regression trees. Toth and Eltinge (2011) evaluated the properties of point estimators with respect to the joint distribution induced by the superpopulation model and the sampling design. In a *iid* setting, Scornet et al. (2015) showed that the population partitions converge to the theoretical partitions. Assumption (H5) can thus be viewed as a design-based version of the result from Scornet et al. (2015). In the Supplementary Material, we conduct a simulation study, whose results suggest that Assumption (H5) seems to be verified, at least in our experiments. More research is needed to provide a rigorous proof of Assumption (H5) in the design-based approach and is beyond the scope of this article.

As in the case of model-assisted estimators based on RF with population-based partitions, we assume that the minimum number of observations, n_{0v}, in the terminal nodes is also growing to infinity and we assume the following additional assumption about the RF resampling algorithm :

(C2) The number of subsampled elements n'_v is such that $\lim_{v\to\infty} n'_v/n_v \in (0;1]$.

This assumption requires that the number n'_v of elements in each subsample increases at the same speed as the sample size n_v, allowing each terminal node to have at least n_{0v} observations.

4.1 Asymptotic results

In this section, we state some results pertaining to sequences of RF model-assisted estimators $\{\widehat{t}_{rf}\}$. The corresponding results for the model-assisted estimators $\{\widehat{t}^*_{rf}\}$ can be found in the Supplementary Material.

Result 4.1. *Consider a sequence of RF model-assisted estimators* $\{\widehat{t}_{rf}\}$. *Then, there*

exist positive constants $\tilde{C}_1, \tilde{C}_2$ *such that*

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf}-t_y\right)\right| \leqslant \frac{\tilde{C}_1}{\sqrt{n_v}}+\frac{\tilde{C}_2}{n_{0v}}, \quad \textit{with } \xi\textit{-probability one.}$$

If $\dfrac{n_v^u}{n_{0v}} = O(1)$ *with* $1/2 \leqslant u \leqslant 1$, *then there exists a positive constant* $\tilde{C}$ *such that*

$$\mathbb{E}_p\left|\frac{1}{N_v}\left(\widehat{t}_{rf}-t_y\right)\right| \leqslant \frac{\tilde{C}}{\sqrt{n_v}}, \quad \textit{with } \xi\textit{-probability one.}$$

Result 4.1 implies that the RF model-assisted estimator $\{\widehat{t}_{rf}\}$ is asymptotically design-unbiased, i.e.,

$$\lim_{v\to\infty} \mathbb{E}_p\left[\frac{1}{N_v}\left(\widehat{t}_{rf}-t_y\right)\right] = 0, \quad \text{with } \xi\text{-probability one,}$$

and design-consistent in the sense that

$$\lim_{v\to\infty} \mathbb{E}_p\left[\mathbf{1}_{\{N_v^{-1}|\widehat{t}_{rf}-t_y|>\eta\}}\right] = 0, \quad \text{with } \xi\text{-probability one}$$

for all $\eta > 0$. Moreover, if n_{0v} is large enough with respect to the sample size n_v, the RF estimator $\widehat{t}_{rf}$ is $\sqrt{n_v}$-consistent. For a given partition, note that the number of terminal nodes is of order $O(n_v/n_{0v})$, and if n_{0v} satisfies the condition from the Result 4.1, the number of terminal nodes is of order $O(n^{1-u})$ for $1/2 \leqslant u \leqslant 1$.

The next result shows that the RF model-assisted estimator $\widehat{t}_{rf}$ is asymptotically equivalent to the pseudo-generalized difference estimator:

$$\widehat{t}_{pgd} = \sum_{k\in U} \widetilde{m}_{rf}(\mathbf{x}_k) + \sum_{k\in S} \frac{y_k - \widetilde{m}_{rf}(\mathbf{x}_k)}{\pi_k}, \tag{22}$$

where $\widetilde{m}_{rf}(\mathbf{x}_k)$ is given by (9).

Result 4.2. *Consider a sequence of RF estimators* $\{\widehat{t}_{rf}\}$. *Assume also that* $\dfrac{n_v^u}{n_{0v}} = O(1)$ *with* $1/2 < u \leqslant 1$. *Then,* $\{\widehat{t}_{rf}\}$ *is asymptotically equivalent to the pseudo-generalized difference estimator* $\widehat{t}_{pgd}$ *in the sense that*

$$\frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{rf}-t_y\right) = \frac{\sqrt{n_v}}{N_v}\left(\widehat{t}_{pgd}-t_y\right) + o_{\mathbb{P}}(1).$$

From Proposition 4.2, it follows that the asymptotic variance of $\widehat{t}_{rf}$ can be approximated by the variance of (22). That is,

$$\mathbb{AV}_p\left(\frac{1}{N_v}\widehat{t}_{rf}\right) \;=\; \mathbb{V}_p\left(\frac{1}{N_v}\widehat{t}_{pgd}\right)$$

$$= \frac{1}{N_v^2} \sum_{k \in U_v} \sum_{\ell \in U_v} (\pi_{kl} - \pi_k \pi_\ell) \frac{y_k - \widetilde{m}_{rf}(\mathbf{x}_k)}{\pi_k} \frac{y_\ell - \widetilde{m}_{rf}(\mathbf{x}_\ell)}{\pi_\ell}. \tag{23}$$

While the RF model-assisted estimator $\widehat{t}_{rf}$ is design-consistent as long as n_{0v} and n_v grow to infinity (Result 4.1), the asymptotic equivalence of $\widehat{t}_{rf}$ with the pseudo-generalized difference estimator $\widehat{t}_{pgd}$ is obtained only for n_{0v} satisfying a certain rate. Stronger assumptions on higher-order inclusion probabilities (Breidt and Opsomer, 2000; McConville and Toth, 2019) are required in order to show that the asymptotic mean squared error of $\widehat{t}_{rf}$ is equivalent to the variance of the pseudo-generalized difference estimator. We do not pursue this further.

Expression (23) suggests that $\widehat{t}_{rf}$ is efficient if the residuals $y_k - \widetilde{m}_{rf}(\mathbf{x}_k)$ are small for all $k \in U_v$. The asymptotic variance given in (23) cannot be computed in practice because the residuals, $y_k - \widetilde{m}_{rf}(\mathbf{x}_k)$, $k \in U$, are unknown. Assuming that $\pi_{k\ell} > 0$ for all pairs $(k, \ell) \in U_v \times U_v$, a design-consistent estimator of $\mathbb{AV}_p\left(\frac{1}{N_v}\widehat{t}_{rf}\right)$ is given by

$$\widehat{\mathbb{V}}_{rf}\left(\frac{1}{N_v}\hat{t}_{rf}\right) = \frac{1}{N_v^2} \sum_{k \in U_v} \sum_{\ell \in U_v} I_k I_\ell \frac{\pi_{k\ell} - \pi_k \pi_\ell}{\pi_{k\ell}} \frac{y_k - \widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k} \frac{y_\ell - \widehat{m}_{rf}(\mathbf{x}_\ell)}{\pi_\ell}, \tag{24}$$

where $\widehat{m}_{rf}(\mathbf{x}_k)$ is given by (17). To establish the design consistency of (24), we require the following additional assumption:

(H6) We assume that $\lim_{v \to \infty} \max_{i,j,k,\ell \in D_{4,N_v}} \left| \mathbb{E}_p \left\{ (I_i I_j - \pi_i \pi_j)(I_k I_\ell - \pi_k \pi_\ell) \right\} \right| = 0$, where D_{4,N_v} denotes the set of distinct 4-tuples from U_v.

Assumption (H6) was suggested by Breidt and Opsomer (2000) and, together with (H2)-(H3), is used to establish the design consistency of the unbiased estimator of the variance of the Horvitz-Thompson estimator $\sum_{k \in S_v} y_k/\pi_k$, assuming that the survey variable Y has finite fourth moment. Assumption (H6) is satisfied for simple random sampling without replacement and stratified simple random sampling without replacement. It is also satisfied for high entropy sampling designs (Boistard et al., 2012; Cardot et al., 2014).

Result 4.3. *Consider a sequence of RF model-assisted estimators* $\{\widehat{t}_{rf}\}$. *Assume also that* $\frac{n_v^u}{n_{0v}} = O(1)$ *with* $1/2 < u \leqslant 1$. *Then, the variance estimator* $\widehat{\mathbb{V}}_{rf}(\widehat{t}_{rf})$ *is asymptotically design-consistent for the asymptotic variance* $\mathbb{AV}_p\left(\widehat{t}_{rf}\right)$. *That is,*

$$\lim_{v \to \infty} \mathbb{E}_p \left(\frac{n_v}{N_v^2} \left| \widehat{\mathbb{V}}_{rf}(\widehat{t}_{rf}) - \mathbb{AV}_p(\widehat{t}_{rf}) \right| \right) = 0.$$

Finally, we establish the central limit theorem that can be used to obtain asymptotically normal confidence intervals of t_y. To that end, we assume that $\widehat{t}_{pgd}$ is normally distributed, an assumption that is satisfied in many classical sampling designs; e.g., see Fuller (2009).

(H7) The sequence of pseudo-generalized difference estimators $\{\widehat{t}_{pgd}\}$ satisfies

$$\frac{N_v^{-1}\left(\widehat{t}_{pgd}-t_y\right)}{\sqrt{\mathbb{V}_p\left(N_v^{-1}\widehat{t}_{pgd}\right)}} \xrightarrow[v\to\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

where $\mathbb{V}_p\left(N_v^{-1}\widehat{t}_{pgd}\right)$ is given by (23).

Result 4.4. *Consider the sequence of RF estimators* $\{\widehat{t}_{rf}\}$. *Then,*

$$\frac{N_v^{-1}\left(\widehat{t}_{rf}-t_y\right)}{\sqrt{\widehat{\mathbb{V}}_{rf}(N_v^{-1}\widehat{t}_{rf})}} \xrightarrow[v\to\infty]{\mathcal{L}} \mathcal{N}\left(0,1\right).$$

The proof of Result 4.4 is a direct application of Results 4.2 and 4.3, and is thus omitted.

5 A model calibration procedure for handling multiple survey variables

In practice, most surveys conducted by national statistical offices (NSO) collect information on multiple survey variables. The collected data are stored in rectangular data files. A column of weights, referred to as a weighting system, is made available on the data file. This weighting system can then be applied to obtain an estimate for any survey variable. However, applying a RF algorithm yield the variable-specific weights (20). In other words, the weights were derived to obtain an estimate of the total for a specific survey variable Y. Hence, applying the weights (20) to other survey variables may produce inefficient estimators. A solution to this issue consists of developing multiple sets of weights, one for each survey variable. This is usually deemed undesirable by data users who are used to work with a single set of weights. In this section, we describe a model calibration procedure (Wu and Sitter, 2001), originally proposed by Montanari and Ranalli (2009), that yields a single weighting system while accounting for multiple survey variables that are deemed important.

Suppose that we can identify a subset of survey variables $Y_1, \ldots, Y_q$, that are deemed important. We postulate the following working model for each variable:

$$\mathbb{E}\left[Y_{jk} \mid \boldsymbol{X}_k = \mathbf{x}_k\right] = m^{(j)}(\mathbf{x}_k^{(j)}), \quad j = 1, \cdots, q, \tag{25}$$

where $m^{(j)}(\cdot)$ is an unknown function and $\mathbf{x}_k^{(j)}$ is a vector of auxiliary variable associated with unit k for the variable Y_j. We allow a different link functions $m(\cdot)$ and different sets of explanatory variables for each of the survey variables $Y_1, \ldots, Y_q$. The interest lies in estimating the population totals $t_{y_1}, \ldots, t_{y_q}$. We assume that each of these totals is estimated using a model-assisted estimator of the form (4) but with possibly different methods. For instance, some of the estimates may be based on a parametric working model, while others may be based on a nonparametric working model (e.g., RF). We can construct the set of q predicted values $\widehat{m}^{(1)}(\mathbf{x}_k^{(1)}), \ldots, \widehat{m}^{(q)}(\mathbf{x}_k^{(q)})$, for $k \in U$.

In addition, we assume that, at the estimation stage, a vector $\mathbf{v}_k$ of size q' of calibration variables is available for $k \in S$ and that the corresponding vector of population totals $\mathbf{t}_{\mathbf{v}} = \sum_{k \in U} \mathbf{v}_k$ is known. In practice, survey managers often want to ensure consistency between survey estimates and known population totals for important variables such as gender and age group.

Given these predictions $\widehat{m}^{(1)}(\mathbf{x}_k^{(1)}), \ldots, \widehat{m}^{(q)}(\mathbf{x}_k^{(q)})$, and the vector calibration variables $\mathbf{v}$, we seek calibrated weights w_k^C, $k \in S$, as close as possible to the initial weights π_k^{-1} subject to the following $q + q' + 1$ calibration constraints:

$$\sum_{k \in S} w_k^C = N, \tag{26}$$

$$\sum_{k \in S} w_k^C \widehat{m}^{(j)}(\mathbf{x}_k^{(j)}) = \sum_{k \in U} \widehat{m}^{(j)}(\mathbf{x}_k^{(j)}), \quad j = 1, \ldots, q, \tag{27}$$

$$\sum_{k \in S} w_k^C \mathbf{v}_k = \sum_{k \in U} \mathbf{v}_k. \tag{28}$$

More specifically, we seek calibrated weights w_k^C such that

$$\sum_{k \in S} G(w_k^C / \pi_k^{-1})$$

is minimized subject to (26)–(28), where $G(\cdot)$ is a pseudo-distance function measuring the closeness between two sets of weights, such that $G(w_k^C/\pi_k^{-1}) \geq 0$, differentiable with respect to w_k^C, strictly convex, with continuous derivatives $g(w_k^C/\pi_k^{-1}) = \partial G(w_k^C/\pi_k^{-1})/\partial w_k^C$ such that $g(1) = 0$; see Deville and Särndal (1992).

The weights w_k^C are given by

$$w_k^C = \pi_k^{-1} F(\widehat{\boldsymbol{\lambda}}^\top \mathbf{h}_k),$$

where $F(.)$ is the calibration function defined as the inverse of $g(.)$, $\widehat{\boldsymbol{\lambda}}$ is a $q + q' + 1$-vector of estimated coefficients and

$$\mathbf{h}_k = \left(1, \widehat{m}_k^{(1)} - \widehat{\overline{m}}^{(1)}, \ldots, \widehat{m}_k^{(q)} - \widehat{\overline{m}}^{(q)}, v_{1k}, \ldots, v_{q'k}\right)^\top \tag{29}$$

with $\widehat{m}_k^{(j)} \equiv m^{(j)}(\mathbf{x}_k^{(j)})$ and $\widehat{\overline{m}}^{(j)} \equiv \sum_{k\in S} \pi_k^{-1} \widehat{m}_k^{(j)} / \sum_{k\in S} \pi_k^{-1}$, $j = 1, \cdots, q$.

The calibrated weights w_k^C may be viewed as a compressed score summarizing the information contained in the q working models (25) and the vector of calibration variables $\mathbf{v}$. The weighting system $\{w_k^C; k \in S\}$ may be then applied to any survey variable Y, which leads to the model calibration type estimator

$$\widehat{t}_{y,mc} = \sum_{k\in S} w_k^C y_k.$$

If the number of calibration constraints $q + q' + 1$ is large, the resulting weights w_k^C may be highly dispersed leading to potentially unstable estimates $\widehat{t}_{y,mc}$. A number of pseudo-distance functions such as the truncated linear and the logit methods may be used to limit the variability of the weights w_k^C; see Deville and Särndal (1992) for a description of these methods. A simple alternative is to use additional constraints on the weights as part of the calibration constraints. For instance, we may impose that $w_k^C < w_0$, where w_0 is a threshold set by the survey statistician; see also Santacatterina and Bottai (2018) for alternative constraints on the weights. Finally, we can relax the calibration constraints (26)-(28) by considering a L^2-penalized criterion, leading to a ridge-type model calibration estimator; see Montanari and Ranalli (2009). Montanari and Ranalli (2009) reports the results of a simulation study, assessing the performance of point estimators obtained through multiple and ridge model calibration methods.

6 Simulation study

6.1 Performance of point estimators

We conducted a simulation study to assess the performance of several model-assisted estimators, in terms of bias and efficiency. We generated a finite population of size

$N = 10,000$, consisting of a set of auxiliary variables and 8 survey variables. We first generated 7 auxiliary variables $X_0, \cdots, X_6$, according to the following distributions: $X_0 \sim \mathcal{U}(0,1)$; $X_1 \sim \mathcal{N}(0,1)$, $X_2 \sim \text{Beta}(3,1)$, $X_3 \sim 2 \times \text{Gamma}(3,2)$, $X_4 \sim \text{Bernoulli}(0.7)$, $X_5 \sim \text{Multinomial}(0.4, 0.3, 0.3)$ and $X_6 \sim \mathcal{E}(1)$. The variables X_1, X_2, X_3, and X_6 have been standardized so as to have a mean and a variance equal to 0 and 1, respectively. To assess the performance of the proposed method in a high-dimensional setting, we also generated 100 additional auxiliary variables $V_1, V_2, \cdots, V_{100}$, from a uniform distribution $\mathcal{U}(-1,1)$. Given the X-variables and the V-variables, we generated the survey variables according to the following models:

Model 1: $Y_1 = 1 + 2(X_0 - 0.5) + \mathcal{N}(0, 0.1)$;

Model 2: $Y_2 = 1 + 2(X_0 - 0.5)^2 + \mathcal{N}(0, 0.1)$;

Model 3: $Y_3 = 2 + X_6 + X_2 + X_3 + X_4 + X_5 + \mathcal{N}(0,1)$;

Model 4: $Y_4 = 2 + (X_6 + X_2 + X_3)^2 + \mathcal{N}(0,1)$;

Model 5: $Y_5 = 0.5X_5 + \exp(-X_1) + 3X_4 + \exp(-X_6) + \mathcal{E}(1)$;

Model 6: $Y_6 = V_1^2 + \exp(-V_2^2) + \mathcal{N}(0, 0.3)$;

Model 7: $Y_7 = V_1^2 + \exp(-V_2^2) + \mathcal{N}(0, 0.3)$;

Model 8: $Y_8 = 3 + V_1V_2 + V_3^2 - V_4V_7 + V_8V_{10} - V_6^2 + \mathcal{N}(0, 0.5)$.

The errors in Model 5 have been scaled and centered so as to have a mean and a variance equal to 0 and 1, respectively. Models 1 and 2 were used in Breidt and Opsomer (2000), while Models 7 and 8 were introduced in Scornet (2017). Models 1-8 were generated so as to include a relatively wide range of relationships between the Y-variable and the set of explanatory variables: linear/non-linear relationships, presence/absence of quadratic terms and presence/absence of interactions. Our scenarios also included low, medium and high-dimensional settings. From the population, we selected $R = 5,000$ samples, of size n, according to simple random sampling without replacement. We used $n = 250$ and $n = 1,000$. In each sample, we computed the following estimators: (i) The Horvitz-Thompson (HT) estimator given by (1); (ii) The generalized regression (GREG) estimator given by (4) with $\widehat{m}(\mathbf{x}_k) = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}$; (iii) The

model-assisted estimator (4) with $\widehat{m}(\mathbf{x}_k)$ obtained through regression trees (CART); and (iv) The model-assisted estimator (4) based on RF, where $\widehat{m}(\mathbf{x}_k)$ is given by (17). We considered three RF algorithms, each based on $1,000$ trees. The first (RF1) was based on bootstrap. The second algorithm (RF2) was based on subsampling with a sampling fraction equal to 0.63 (Scornet, 2017). For both RF1 and RF2, the minimum number of observations per terminal node was set to $n_0 = 5$. Finally, the third algorithm (RF3) was based on bootstrap with $n_0 = \sqrt{n}$ observations in each terminal node. In RF1-RF3, we used $m_{try} = \sqrt{p}$ as it is the default number of variables considered for the splitting process in most software packages dealing with RF for regression.

For the estimators GREG, CART, RF1, RF2 and RF3, the predictions $\widehat{m}(\mathbf{x}_k)$ were obtained using the working models described in Table 1. For the survey variables Y_7 and Y_8, the working models were based on a large number of superfluous explanatory variables (50 and 100, respectively), which allowed us to assess the behavior of the resulting estimators in a medium/high dimensional setting.

Table 1: The working models

Survey variable	Vector of explanatory variable $\mathbf{X}$ used in the working model
Y_1	X_0
Y_2	X_0
Y_3	$X_1 - X_6$
Y_4	$X_1 - X_6$
Y_5	$X_1 - X_6$
Y_6	$V_1 - V_{10}$
Y_7	$V_1 - V_{50}$
Y_8	$V_1 - V_{100}$

We were interested in estimating the population totals $t_{y_j} = \sum_{k \in U} y_{kj}$, $j = 1, \ldots, 8$. As a measure of bias of an estimator $\hat{t}_{y_j}$, we used the Monte Carlo percent relative bias defined as

$$RB(\widehat{t}_{y_j}) = 100 \times \frac{1}{R} \sum_{r=1}^{R} \frac{(\widehat{t}_{y_j}^{(r)} - t_{y_j})}{t_{y_j}},$$

where $\widehat{t}_{y_j}^{(r)}$ denotes the estimator $\widehat{t}_{y_j}$ in the rth iteration, $r = 1, ..., R$. As a measure of efficiency of an estimator $\hat{t}_{y_j}$, we used the relative efficiency, using the Horvitz-Thompson estimator, $\widehat{t}_{y_j,\pi}$, as the reference:

$$RE(\widehat{t}_{y_j}) = 100 \times \frac{MSE(\widehat{t}_{y_j})}{MSE(\widehat{t}_{y_j,\pi})},$$

where

$$MSE(\widehat{t}_{y_j}) = \frac{1}{R}\sum_{r=1}^{R}(\widehat{t}_{y_j}^{(r)} - t_{y_j})^2$$

and $MSE(\widehat{t}_{y_j,\pi})$ is defined similarly. The results are displayed in Tables 2 and 3. The simulations were performed using the R software with the package *ranger* (Wright and Ziegler, 2015).

Table 2: Monte Carlo percent relative bias (RB) and Monte Carlo efficiency (RE) of several model-assisted estimators for $n = 250$

Population		GREG	CART	RF1	RF2	RF3
Y_1	RB	-0.0	-0.0	-0.0	-0.0	0.0
	RE	3.0	3.5	3.7	3.6	3.4
Y_2	RB	-0.0	0.0	0.0	0.0	0.0
	RE	101.0	37.6	39.4	38.3	35.0
Y_3	RB	0.0	-0.0	-0.1	-0.1	-0.0
	RE	19.6	55.2	33.8	34.0	35.4
Y_4	RB	-0.7	-1.2	-1.2	-1.5	-0.7
	RE	81.1	61.1	49.7	49.0	53.1
Y_5	RB	-0.1	0.1	-0.0	-0.0	-0.0
	RE	37.9	32.7	25.8	26.5	30.7
Y_6	RB	-0.0	0.3	-0.0	-0.0	-0.0
	RE	105.2	72.2	57.5	57.5	58.3
Y_7	RB	-0.0	0.2	0.1	0.0	0.0
	RE	127.6	84.3	75.8	75.5	76.8
Y_8	RB	0.0	0.0	0.0	0.0	0.0
	RE	127.0	135.6	92.7	92.5	95.6

We start by noting that all the estimators displayed a negligible bias in all the scenarios, as expected. Also, both RF1 and RF2 showed very similar performances in terms of bias and efficiency in all the scenarios. This is consistent with the empirical results of Scornet (2017); i.e., the strategy based on bootstrap and the strategy based on subsampling with a sampling fraction of 0.63 led to similar performances. The results for RF3 were similar to those obtained for RF1 and RF2, which suggests that the number of observations in each terminal node did not seem to affect the behavior of the point estimator, at least in our experiments. This may not be the case in other scenarios as we illustrate in Section 6.3.

Table 3: Monte Carlo percent relative bias (RB) and Monte Carlo efficiency (RE) of several model-assisted estimators for $n = 1000$.

Population		GREG	CART	RF1	RF2	RF3
Y_1	RB	0.0	0.0	0.0	0.0	0.0
	RE	2.8	3.5	3.6	3.5	3.0
Y_2	RB	0.0	0.0	0.0	0.0	0.0
	RE	100.1	38.7	40.5	39.6	33.3
Y_3	RB	0.0	0.0	-0.1	-0.1	0.0
	RE	20.4	41.1	28.1	27.8	31.6
Y_4	RB	-0.1	-1.1	-0.9	-0.7	-0.2
	RE	78.9	52.3	36.7	36.1	44.5
Y_5	RB	-0.0	0.0	0.0	0.0	-0.0
	RE	37.3	24.5	20.9	21.2	24.8
Y_6	RB	0.0	0.0	-0.0	-0.0	-0.0
	RE	101.1	65.5	49.1	49.2	50.3
Y_7	RB	0.0	0.0	0.0	0.0	0.0
	RE	105.5	73.2	63.3	63.2	65.0
Y_8	RB	-0.0	-0.0	-0.0	-0.0	0.0
	RE	166.6	137.6	96.0	95.7	89.5

In the case of a linear relationship (which corresponds to the survey variables Y_1 and Y_3), the GREG estimator was the most efficient, as expected. For instance, for the survey variables Y_3, the value of RE for the GREG estimator was about 19.6%, whereas the RF1, RF2 and RF3 estimators showed a value of RE of about 34%. In the case of a nonlinear relationship (which corresponds to the survey variables Y_2 and $Y_4, \ldots, Y_8$), the GREG estimator was less efficient than RF1, RF2 and RF3. For instance, in the case of the variable Y_4, the GREG showed a value of RE of about 81.1%, whereas the RE of RF estimators lied between 49.0% and 53.1%. For the variables Y_6, Y_7, Y_8, the GREG estimator was even less efficient than the Horvitz-Thompson estimator with values of RE ranging from 105% to 127%.

In the case of a single explanatory variable (which corresponds to the survey variables Y_1 and Y_2), RF and regression trees displayed very similar performances. In contrast, the estimators RF1, RF2 and RF3 were more efficient than the CART esti-

mator when the vector of explanatory variables was multi-dimensional (i.e., variables $Y_3, \ldots, Y_9$). In a high-dimensional setting (which corresponds to the survey variables Y_7 and Y_8), the RF estimators were more efficient than the Horvitz-Thompson estimator, even for $n = 250$.

6.2 Performance of the proposed variance estimator

We have also investigated the performance of the variance estimator $\widehat{\mathbb{V}}_{rf}$ given by (24) in the case of RF with subsampling, in terms of relative bias and coverage of normal-based confidence intervals. We generated a population of size $N = 100,000$ according to Model 5. The sample size was set to $n = 500; 1,000; 5,000; 10,000; 20,000$ and $50,000$. Here, we present the results for $B = 1$ but other values of B led to similar results and are not shown here. As we suspected that the number of observations in each terminal node, n_0, may have an impact on the behavior of $\widehat{\mathbb{V}}_{rf}$, we used different values for n_0 : $n_0 = \lfloor n^{a/20} \rfloor$ for $a = 1; 3; 5; 7; 9; 11; 13; 15; 17$. The choice $n_0 = \lfloor n^{11/20} \rfloor$ was advocated by McConville and Toth (2019). Figure 2 shows the Monte Carlo percent relative bias of $\widehat{\mathbb{V}}_{rf}$ for different values of n and n_0. Figure 3 shows the Monte Carlo coverage rate of the confidence interval, $\widehat{t}_{rf} \pm 1.96\sqrt{\widehat{\mathbb{V}}_{rf}}$, for different values of n and n_0.

From Figure 2, we note that $\widehat{\mathbb{V}}_{rf}$ is severely biased for small values of n_0 and as a consequence, the confidence intervals (see Figure 3) perform poorly for small values of n_0 because of the substantial underestimation of the true variance in these scenarios. For a given value of n_0, we note that the bias decreases as n increases and for a given value of n, the bias decreases as n_0 increases. For $n_0 = \lfloor 13/20 \rfloor$, the confidence intervals perform relatively well with coverage rates close to the nominal rate. The significant bias for small values of n_0 is most likely due to overfitting, which is characterized by the presence of artificially small residuals $y_k - \widehat{m}(\mathbf{x}_k)$ in each terminal node, which in turn, leads to underestimation. This issue was raised by Opsomer and Miller (2005) in the context of local polynomial regression. To cope with this issue, we suggest a variance estimator based on a K-fold criterion. More specifically, we randomly split the sample S into K groups $S_\kappa, \kappa = 1, \ldots, K$, of approximately equal size. For $k \in S_\kappa$, let $\widehat{m}^{(-\kappa)}(\mathbf{x}_k)$ denote the prediction at the point $\mathbf{x}_k$ built on $S - S_\kappa$ and $\widehat{\epsilon}_k^{(-\kappa)} = y_k - \widehat{m}^{(-\kappa)}(\mathbf{x}_k)$ the associated residual. The proposed K-fold variance estimator is given by

$\widehat{V}^{(K)} = \sum_{\kappa_1=1}^{K} \sum_{\kappa_2=1}^{K} \sum_{k \in \mathcal{S}_{\kappa_1}} \sum_{\ell \in \mathcal{S}_{\kappa_2}} (\Delta_{k\ell}/\pi_{k\ell})(\widehat{\epsilon}_k^{(-\kappa_1)}/\pi_k)(\widehat{\epsilon}_\ell^{(-\kappa_2)}/\pi_\ell)$. In practice, the number of groups (or folds) is often set to $K = 5$ or $K = 10$. We tested the performance of $\widehat{V}^{(5)}$ in terms of bias and coverage probability, using the same scenarios as above. The bias was almost negligible for all sizes n and n_0 and the coverage rates lied between 93% and 96%, which constitutes a significant improvement over the results displayed in Figures 2 and 3. More research is needed in order to establish the theoretical properties of the variance estimator based on a K-fold criterion evaluate. It will be treated elsewhere.

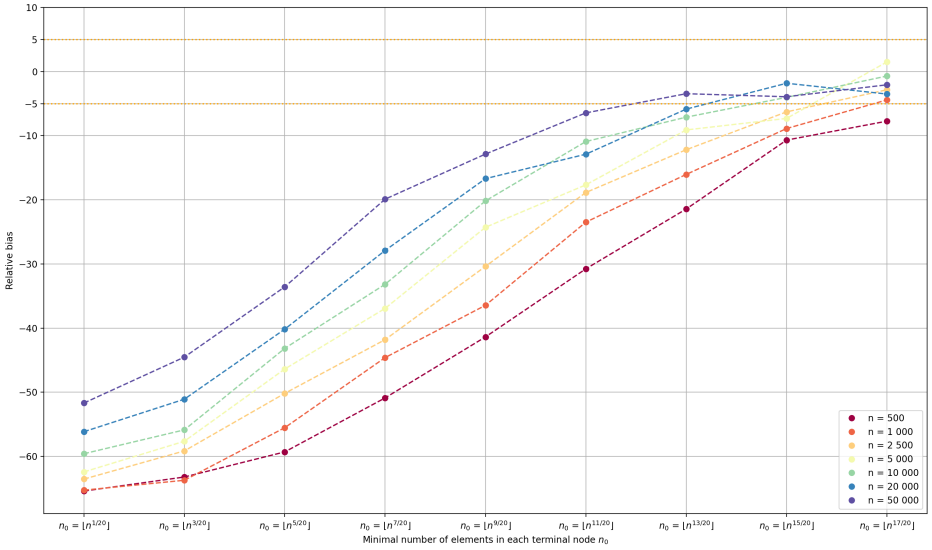

Figure 2: Evolution of the relative bias with respect to n_0.

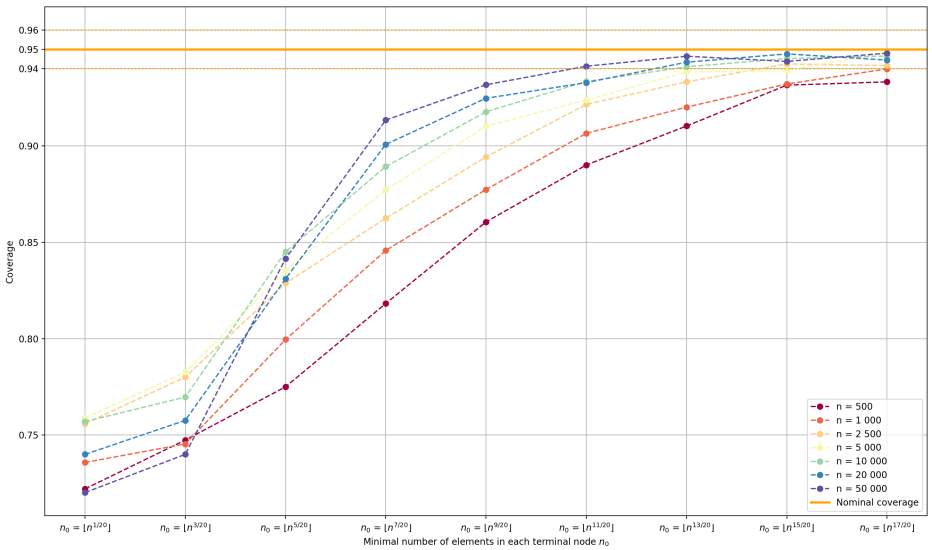

Figure 3: Evolution of the effective coverage with respect to n_0.

6.3 Choice of hyper-parameters

To get a better understanding of how the choice of hyper-parameters impacts the behavior of model-assisted estimators based on RF, we conducted additional scenarios. We first identified the following important hyper-parameters involved in the RF algorithm of Breiman (2001):

i) The minimal number of observations, n_0, in each terminal node;

ii) The number of trees in the forest B;

iii) The number of variables considered for the search of the best split in the optimization criterion (16);

iv) The resampling mechanism.

The additional scenarios were conducted using a finite population of size $N = 10,000$ consisting of the survey variables Y_5 and Y_8 described in Section 6.1. Recall that the working model for the survey variables Y_5 included the predictors $X_1, \ldots, X_6$, whereas it included the predictors $V_1, \ldots, V_{100}$ for the variable Y_8 (see Table 1).

From the population, we generated $R = 10,000$ samples, of size $n = 1,000$, according to simple random sampling without replacement. Figure 4 and Figure 5 show, respectively, the relative efficiency of the model-assisted estimators based on RF, $\widehat{t}_{rf}$, corresponding to Y_5 and Y_8, respectively, for several values of n_0. Figure 4 suggests that $\widehat{t}_{rf}$ was much more efficient than the Horvitz-Thompson estimator for small values of n_0 and that the value of RE approached 100 as n_0 increased. This result can be explained by the fact that small values of n_0 led to homogeneous terminal nodes, which in turn led to small residuals $y_k - \widehat{m}_{rf}(\mathbf{x}_k)$. For the survey variable Y_8, we note from Figure 5 that the value of n_0 did not seem to affect the efficiency of the corresponding model-assisted estimator.

Figure 6 display the relative efficiency for several values of B, the number of trees in the forest for the survey variable Y_8. As expected, a small value of B causes the estimator $\widehat{t}_{rf}$ to loose some efficiency. Figure 6 suggests that $B = 50$ led to good results and that the efficiency of $\widehat{t}_{rf}$ was not much affected by the number of trees B for $B \geq 50$. Nevertheless, it is advisable to choose a large value of B if the computational

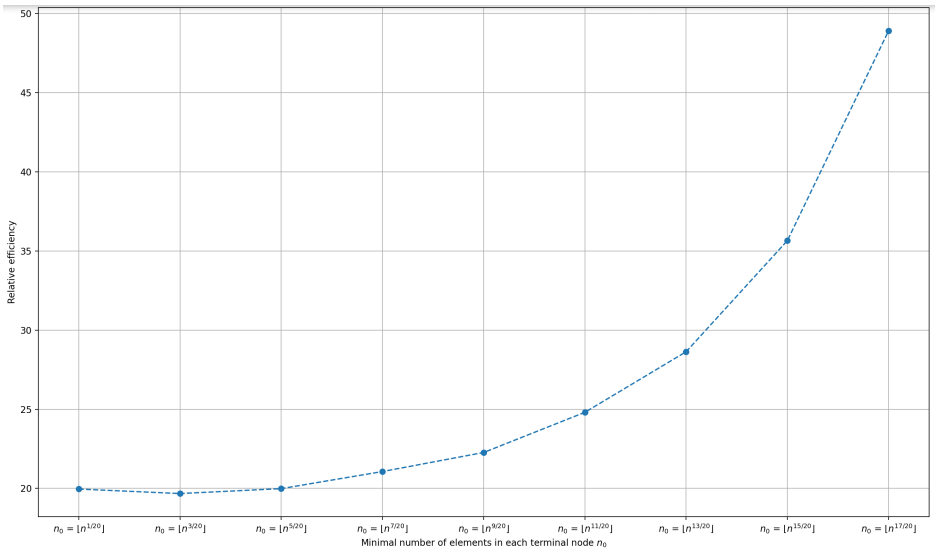

Figure 4: Relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_5 and for several values of n_0.

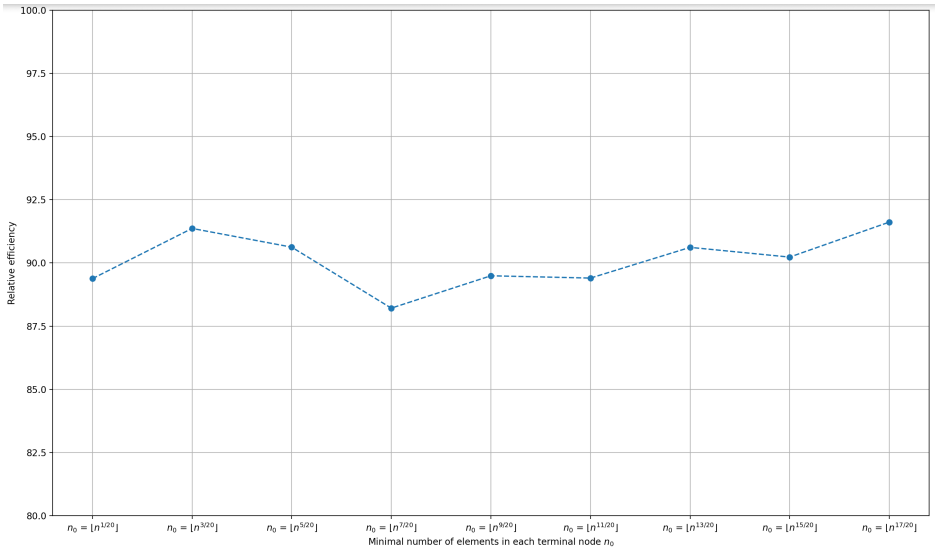

Figure 5: Relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_8 and for several values of n_0

capacity permits. The results for the survey variable Y_5 were very similar and so we omit them.

In most software packages, the default number of variables considered for the splitting process is $m_{try} = \sqrt{p}$ in case of regression. In our simulations, this choice led to satisfactory results in most scenarios. Figure 7 shows the relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_8 and for several values of m_{try}. Since the working model for Y_8 contained $p = 100$ explanatory variables, the default value $\sqrt{p}$ was equal to 10.

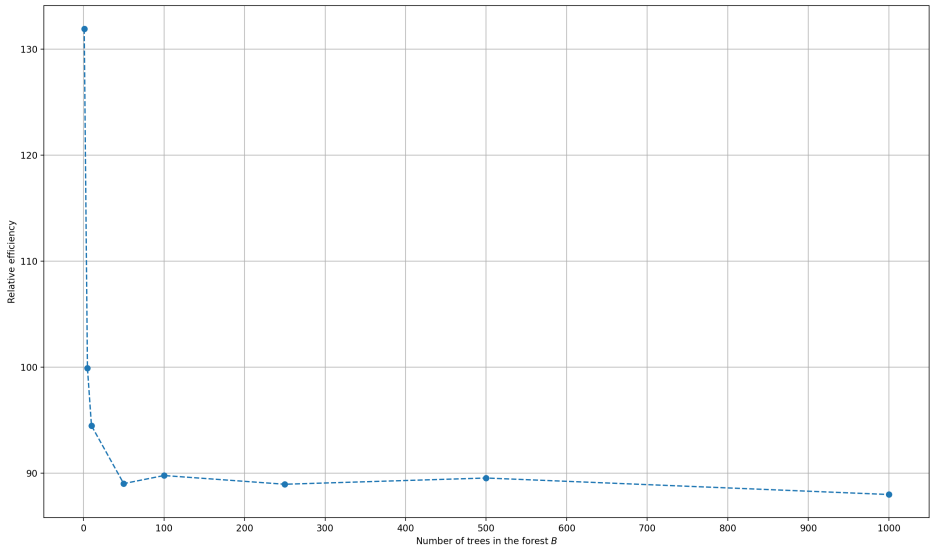

Figure 6: Relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_8 and for several values of B.

Although the value $\sqrt{p} = 10$ was not the best choice for optimal performances, it led efficient model-assisted estimators. Furthermore, the relative efficiency did not vary much for values B larger than 30.

Turning to the resampling mechanism, a common choice is to use bootstrap (with

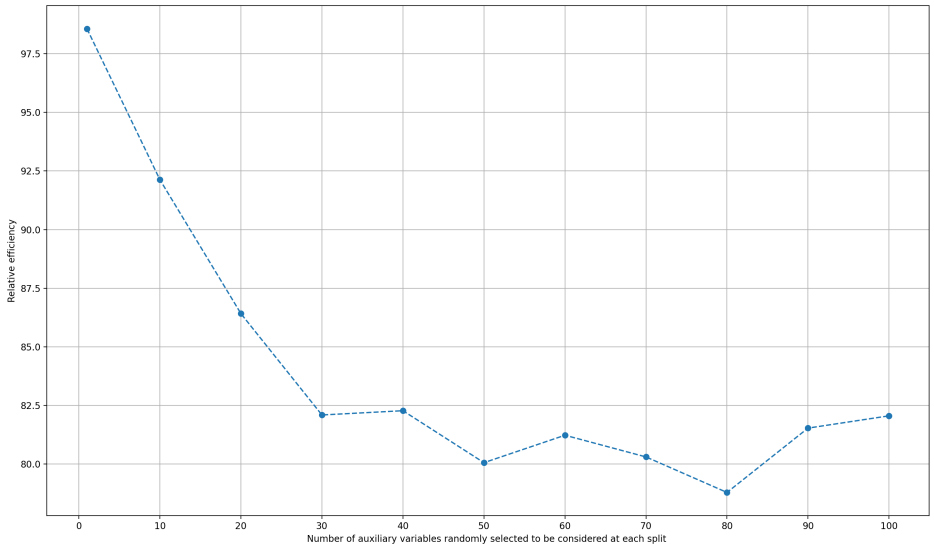

Figure 7: Relative efficiency of $\widehat{t}_{rf}$ for the survey variable Y_8 and for several values of m_{try}.

replacement), for which some of the results presented in the paper do not apply. However, as noted by several authors (see e.g. Scornet et al. (2015), Wager (2014) and the references therein) and as shown in our simulations, selecting the points without

replacement rather than with replacement does not seem to affect the performance of the resulting model-assisted estimators in most cases.

6.4 Real data application

In this section, we apply the proposed methods using data collected by Médiamétrie, the company that measures the media audience in France. In this application, we focus on radio audiences. Each year, Médiamétrie conducts a survey aiming at gathering detailed information about French individuals 13 years of age and over, including socio-demographic variables and radio listening habits. We used the 2019 radio audience data that consisted of $N = 26,293$ individuals. As a survey variable, we considered the binary variable Y, such that $y_k = 1$ if an individual in the kth individual listens to the radio of interest on a daily basis, and $y_k = 0$ otherwise. For confidentiality reasons, we omit the name of the radio broadcaster. We aimed at estimating the proportion of French individuals who listen to the radio of interest on a daily basis, both at the overall population level and for several domains of interest. For each individual, we had access to 43 socio-demographic variables (e.g., number of individuals in the household, age of each member of the household, gender, internet habits, occupation, etc.) and their listening habits of 21 other radios. For each individual, we also knew whether or not the individual listens to any of these 21 radios, for each interval of 7.5 minutes on a typical day. This led to a data set with $p = 3,882$ variables, among which $3,839$ were binary.

From the data set, we selected a single sample of size $n = 4,000$ according to a stratified sampling design with 5 strata, each stratum corresponding to a French region: North-East, North-West, Île-de-France, South-east and South-West. The strata sample sizes were determined according to proportional allocation. We considered the following domains of interest: the sub-population of individuals who connects to the internet everyday, almost every day, once or twice per week, once to three times per month, very rarely, never, the sub-population of individuals with/without children, the sub-populations of individuals living in cities of size (less than $20,000$, between $20,000$ and $50,000$, between $50,000$ and $100,000$, between $100,000$ and $200,000$ and larger than $200,000$) and the sub-population of individuals living in households of size $1, 2, 3, 4, 5$ and $5+$

We computed the following estimates both at the overall level and at the domain level: (i) The Horvitz-Thompson estimator; (ii) the GREG estimator and (iii) the model-assisted estimator based on RF with hyper-parameters $B = 1,000$, $n_0 = \lfloor n^{11/20} \rfloor$ and $m_{try} = \sqrt{p}$. The working models used for the GREG estimator and the model-assisted RF estimator included 3882 explanatory variables. In each scenario, we also computed a 95% confidence interval for the proportion in the population of individuals who listen to the radio of interest. Finally, we computed the ratio of the estimated variances, using the estimated variance of the Horvitz-Thompson estimator, as the reference. Note that the "true value" was known for each domain of interest. The results (in percentage) are given in Table 4.

From Table 4, we note that the Horvitz-Thompson estimator performed relatively well in most scenarios. Because of the large number of predictors, the GREG estimator suffered from significant small sample bias. For instance, the estimate based on the GREG estimate at the overall level was equal to 27.7%, far from the true value of about 13.5%. In terms of point estimation, the RF model-assisted estimator led to very similar results than those obtained with the Horvitz-Thompson estimator. However, RF led to substantial improvement in terms of estimated variance. Indeed, out of the 22 domains, the value of RV(RF) was smaller than 0.65 for 20 domains. The results suggest that, unlike the GREG estimator, the model-assisted estimator based on RF was not affected by the large number of explanatory variables in the working model. The median length of the confidence intervals was equal to 5.4% for the Horvitz-Thompson estimator, 4.2% for the RF estimator and 3.8% for the GREG estimator.

Domain	True value	HT	RF	GREG	RV (RF)	RV (GREG)	CI HT	CI RF	CI GREG
Overall	13.5	13.7	13.6	27.7	64.6	48.5	[12.7; 14.7]	[12.8; 14.4]	[27.0; 28.3]
Freq: Every day	14.6	14.6	14.4	28.1	61.3	44.0	[13.5; 15.8]	[13.5; 15.3]	[27.3; 28.9]
Freq: Almost every day	14.0	15.5	14.9	31.8	62.0	52.7	[11.1; 19.8]	[11.4; 18.3]	[28.6; 35.0]
Freq: 1-2 / week	8.7	10.5	11.2	55.1	67.0	62.9	[6.1; 14.9]	[7.6; 14.7]	[51.6; 58.5]
Freq: 1-3 / month	11.3	16.1	12.8	3.8	47.6	28.1	[5.8; 26.4]	[5.6; 19.9]	[-1.5; 9.3]
Freq: Very rarely	6.7	7.6	10.0	29.6	60.5	57.7	[1.4; 13.8]	[5.2; 14.8]	[24.9; 34.3]
Freq: Never	8.4	2.5	8.1	-66.1	103	125	[-2.0; 7.0]	[3.4; 12.7]	[-71.2; -61.1]
Children: Yes	11.1	11.5	11.4	33.0	57.2	40.9	[9.7; 13.3]	[10.0; 12.7]	[31.9; 34.2]
Children: No	14.4	14.5	14.5	25.7	62.3	47.4	[13.3; 15.7]	[13.5; 15.4]	[24.8; 26.5]
Inhabitants: Country	12.7	12.8	12.6	45.4	58.9	42.6	[10.5; 15.1]	[10.8; 14.3]	[44.0; 46.9]
Inhabitants: <20K	12.9	12.7	12.1	35.2	52.6	37.2	[10.1; 15.3]	[10.2; 14.0]	[33.6; 36.8]
Inhabitants: 20-50K	12.5	10.8	11.0	36.9	62.2	52.4	[6.7; 14.9]	[7.8; 14.3]	[33.9; 39.8]
Inhabitants: 50-100K	12.3	11.5	12.2	25.9	58.6	47.7	[8.8; 14.2]	[10.1; 14.3]	[24.0; 27.8]
Inhabitants: 100-200K	15.5	18.0	17.9	5.8	53.4	38.5	[14.0; 22.0]	[15.0; 20.8]	[3.3; 8.2]
Inhabitants: >200K	14.2	14.0	14.0	27.3	60.4	46.1	[12.1; 16.0]	[12.5; 15.6]	[26.0; 28.7]
Inhabitants: Paris	14.3	16.3	16.0	0.6	60.7	45.3	[13.0; 19.6]	[13.4; 18.6]	[-1.5; 2.8]
N Household: 1	13.9	13.1	13.5	8.2	58.5	44.8	[11.1; 15.1]	[12.0; 15.1]	[6.8; 9.5]
N Household: 2	16.3	15.9	16.1	27.3	58.1	45.7	[14.0; 17.8]	[14.6; 17.5]	[26.1; 28.6]
N Household: 3	11.7	14.2	13.4	33.9	55.3	37.3	[11.4; 15.4]	[11.5; 16.9]	[32.3; 35.6]
N Household: 4	11.3	12.7	11.8	50.3	56.5	41.0	[10.2; 15.2]	[9.9; 13.7]	[48.7; 51.9]
N Household: 5	9.7	8.2	9,1	28.4	63.1	46.2	[4.9; 11.6]	[6.5; 11.8]	[26.1; 30.6]
N Household: >5	6.3	5.0	3.9	50.9	53.8	26.8	[0.0; 9.5]	[0.0; 7.2]	[48.5; 53.2]

Table 4: Point and relative variance estimates for the percentage of household who listen to the radio of interest for twenty-two domain of interest and associated 95% confidence intervals

7 Final remarks

In this paper, we have introduced a new class of model-assisted estimators based on random forests and derived corresponding variance estimators. We have established the theoretical properties of point and variance estimators obtained through a RF algorithm based on subsampling. The results of an empirical study suggest that the proposed estimators perform well in a wide variety of settings, unlike the GREG and CART estimators. In practice, this robustness property is especially attractive when the data and the underlying relationships are complex. The application on radio audience data recorded by the French company Médiamétrie showed that the RF proposed estimator performed well in this high-dimension setting. We have also described a model calibration procedure for handling multiple survey variables, yet producing a single set of weights, which is attractive from a data user's perspective.

In practice, virtually all survey face the problem of missing values. Survey statisticians distinguish unit nonresponse (when no information is collected on a sampled unit) from item nonresponse (when the absence of information is limited to some variables only). The treatment of unit nonresponse starts with postulating a nonresponse model describing the relationship between the response indicators (equal to 1 for respondents and 0 for nonrespondents) and a vector of explanatory variables. The treatment of item nonresponse starts with postulating an imputation model describing the relationship between the variable requiring imputation and a set of explanatory variables. In both unit and item nonresponse, determining a suitable model is crucial. Therefore, regression trees and RF may prove useful for obtaining accurate estimated response propensities and predicted values. To the best of our knowledge, a theoretical treatment of regression trees and RF in the context of either unit nonresponse or item nonresponse in a finite population setting is lacking. These topics are currently under investigation.

Traditionally, survey samples have been collected through probability sampling procedures and inferences were conducted with respect to the customary design-based framework. In recent years, there has been a shift of paradigm that can be explained by three main factors: (i) a dramatic decrease of response rates; (ii) a rapid increase in data collection costs; and (iii) the proliferation of nonprobabilistic data sources (e.g.,

administrative files, web survey panels, social media data, satellite information, etc.). To meet these new challenges, survey statisticians face increasing pressure to utilize these convenient but often uncontrolled data sources. While such sources provide timely data for a large number of variables and population elements, they often fail to represent the target population of interest because of inherent selection biases. The integration of data from a nonprobability source with data from a probability survey is a topic that is currently being scrutinized by National Statistical Offices. An approach to data integration is statistical matching or mass imputation; see Yang and Kim (2020) for a very recent review on the topic. Again, regression trees and RF algorithms may prove useful in the context of integration of survey data. This topic is currently under investigation.

In a high-dimensional setting, RF may be used to select the most predictive predictors, which in turn may be used in the construction of model-assisted estimators of population totals/means. In this context, issues such as variable selection bias (Strobl et al., 2007) in a finite population setting need to be investigated. This will be treated elsewhere.

Appendix

Proof of Proposition 2.1 Since

$$\widetilde{W}_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}}\left(\mathbf{x}_k, \theta_b^{(U)}\right)}{\widetilde{N}(\mathbf{x}_k, \theta_b^{(U)})}, \tag{30}$$

involves positive quantities only, the weights $\widetilde{W}_\ell(\mathbf{x}_k)$ are nonnegative. Since $\psi_\ell^{(b,U)} \in \{0,1\}$ for all $\ell \in U$ and for all $b \in 1, 2, \ldots, B$, the weight can be bounded as follows:

$$\begin{aligned}\widetilde{W}_\ell(\mathbf{x}_k) = \frac{1}{B}\sum_{b=1}^{B} \frac{\psi_\ell^{(b,U)} \mathbb{1}_{\mathbf{x}_\ell \in A^{(U)}}\left(\mathbf{x}_k, \theta_b^{(U)}\right)}{\widetilde{N}(\mathbf{x}_k, \theta_b^{(U)})} &\leqslant \frac{1}{B}\sum_{b=1}^{B} \left(\widetilde{N}\left(\mathbf{x}_k, \theta_b^{(U)}\right)\right)^{-1} \\ &\leqslant cN_0^{-1}.\end{aligned}$$

where c does not depend on b nor on k or ℓ. To show ii), fix $b \in 1, 2, ..., B$. The result follows by noting that $\widetilde{W}_\ell(\mathbf{x}_k) = \left(\widetilde{N}\left(\mathbf{x}_k, \theta_b^{(U)}\right)\right)^{-1}$ exactly $\widetilde{N}\left(\mathbf{x}_k, \theta_b^{(U)}\right)$ times.

Proof of Proposition 3.2 Let $\{\widehat{t}^{(b)}\}$ be a sequence of estimators of t_y. Then,

$$\mathbb{V}_p\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{t}^{(b)}\right) = \frac{1}{B^2}\sum_{b=1}^{B}\left(\mathbb{V}_p\left(\widehat{t}^{(b)}\right) + \sum_{b=1}^{B}\sum_{b\neq b'=1}^{B} \mathbb{C}or_p\left(\widehat{t}^{(b)}, \widehat{t}^{(b')}\right) \mathbb{V}_p^{1/2}\left(\widehat{t}^{(b)}\right) \mathbb{V}_p^{1/2}\left(\widehat{t}^{(b')}\right)\right).$$

$$\leqslant \frac{\mathbb{V}_p(\widehat{t}^{(1)})}{B} + \mathbb{V}_p(\widehat{t}^{(1)}) \max_{b \neq b'} \left| \mathbb{C}or_p\left(\widehat{t}^{(b)}, \widehat{t}^{(b')}\right)\right|$$

and $\text{Bias}_p^2(B^{-1}\sum_{b=1}^B \widehat{t}^{(b)}) = \text{Bias}_p^2(\widehat{t}^{(1)}) = MSE_p(\widehat{t}^{(1)}) - \mathbb{V}_p(\widehat{t}^{(1)})$. So, for B large enough:

$$MSE_p\left(\frac{1}{B}\sum_{b=1}^B \widehat{t}^{(b)}\right) \leqslant \mathbb{V}_p(\widehat{t}^{(1)}) \max_{b \neq b'} \left| \mathbb{C}or_p\left(\widehat{t}^{(b)}, \widehat{t}^{(b')}\right)\right| - \mathbb{V}_p(\widehat{t}^{(1)}) + MSE_p(\widehat{t}^{(1)}).$$

Proof of Proposition 3.3 Consider the B partitions build at the sample level $\widehat{\mathcal{P}}_S = \{\widehat{\mathcal{P}}_S^{(b)}\}_{b=1}^B$. For a given $b = 1, \ldots, B$, the partition $\widehat{\mathcal{P}}_S^{(b)}$ is composed by disjointed regions as follows $\widehat{\mathcal{P}}_S^{(b)} = \{A_j^{(bS)}\}_{j=1}^{J_{bS}}$ and for each b, consider the J_{bS} dimensional vector $\hat{\mathbf{z}}_k^{(b)} = \left(\mathbb{1}_{\mathbf{x}_k \in A_1^{(bS)}}, \ldots, \mathbb{1}_{\mathbf{x}_k \in A_{J_{bS}}^{(bS)}}\right)^\top$ where $\mathbb{1}_{\mathbf{x}_k \in A_j^{(bS)}} = 1$ if $\mathbf{x}_k$ belongs to the region $A_j^{(bS)}$ and zero otherwise for all $j = 1, \ldots, J_{bS}$. Since $\{A_j^{(bS)}\}_{j=1}^{J_{bS}}$ is a partition, then $\mathbf{x}_k$ will belong to only one region and so, the vector $\hat{\mathbf{z}}_k^{(b)}$ will contain only one non zero component. We have $\widehat{m}_{rf}(\mathbf{x}_k) = B^{-1}\sum_{b=1}^B \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})$ and $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})$ can be written as $\widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}) = (\hat{\mathbf{z}}_k^{(b)})^\top \widehat{\boldsymbol{\beta}}^{(b)}$ where $\widehat{\boldsymbol{\beta}}^{(b)} = \left(\sum_{\ell \in S} \pi_\ell^{-1} \psi_\ell^{(b,S)} \hat{\mathbf{z}}_\ell^{(b)} (\hat{\mathbf{z}}_\ell^{(b)})^\top\right)^{-1} \sum_{\ell \in S} \pi_\ell^{-1} \psi_\ell^{(b,S)} \hat{\mathbf{z}}_\ell^{(b)} y_\ell$ (see also the supplementary materiel for more details). Now,

$$\sum_{k \in S} \frac{y_k - \widehat{m}_{rf}(\mathbf{x}_k)}{\pi_k} = \frac{1}{B}\sum_{b=1}^B \sum_{k \in S} \frac{y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)})}{\pi_k} = \frac{1}{B}\sum_{b=1}^B \sum_{k \in S} \frac{(1 - \psi_k^{(b,S)})(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}))}{\pi_k} + \frac{1}{B}\sum_{b=1}^B \sum_{k \in S} \frac{\psi_k^{(b,S)}(y_k - \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}))}{\pi_k}.$$

For each b, consider the J_{bS} dimensional vector $\mathbf{1}_{J_{bS}}$ whose elements are all equal to one, we have then $\mathbf{1}_{J_{bS}}^\top \hat{\mathbf{z}}_k^{(b)} = 1$ for all k, so

$$\sum_{k \in S} \frac{\psi_k^{(b,S)}}{\pi_k} \widehat{m}_{tree}^{(b)}(\mathbf{x}_k, \theta_b^{(S)}) = \sum_{k \in S} \frac{\psi_k^{(b,S)}}{\pi_k} (\hat{\mathbf{z}}_k^{(b)})^\top \widehat{\boldsymbol{\beta}}^{(b)} = \mathbf{1}_{J_{bS}}^\top \sum_{k \in S} \frac{\psi_k^{(b,S)}}{\pi_k} \hat{\mathbf{z}}_k^{(b)} (\hat{\mathbf{z}}_k^{(b)})^\top \widehat{\boldsymbol{\beta}}^{(b)} = \sum_{\ell \in S} \frac{\psi_\ell^{(b,S)}}{\pi_\ell} y_\ell.$$

References

Arnould, L., Boyer, C., and Scornet, E. (2020). Analyzing the tree-layer structure of deep forests. *arXiv preprint arXiv:2010.15690*.

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24:2546–2554.

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

Biau, G. and Devroye, L. (2014). Cecllular tree classifiers. In *International Conference on Algorithmic Learning Theory*. Springer.

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.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25(2):197–227.

Boistard, H., Lopuhaä, H. P., and Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electron. J. Stat.*, 6:1967–1983.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92:831–846.

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

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205.

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 & Hall/CRC, Boca Raton.

Buskirk, T. D. and Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field*, pages 1–17.

Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2010). Properties of design-based functional principal components analysis. *J. of Statistical Planning and Inference*, 140:75–91.

Cardot, H., Goga, C., and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7:562–596.

Cardot, H., Goga, C., and Lardin, P. (2014). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. *Scandinavian J. of Statistics*, 41:516–534.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

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

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Devroye, L., Gyorfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

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

Firth, D. and Bennett, K. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):3–21.

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.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Fuller, W.-A. (2009). *Sampling Statistics*. John Wiley & Sons.

Genuer, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression. *Canad. J. Statist.*, 33(2):163–180.

Goga, C. and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society, B*, 76:113–140.

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.

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.

Han, T., Jiang, D., Zhao, Q., Wang, L., and Yin, K. (2018). Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. *Transactions of the Institute of Measurement and Control*, 40(8):2681–2693.

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

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

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2015). *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics.

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).

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. In *Survey Research Methods*, volume 13, pages 73–93.

Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24:51–56.

McConville, K. and Breidt, F. J. (2013). Survey design asymptotics for the model-assisted penalised spline regression estimator. *Journal of Nonparametric Regression*, 25(3):745–763.

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.

Montanari, G. and Ranalli, M. G. (2009). Multiple and ridge model calibration for sample surveys. unpublished.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration in survey sampling. *J. Amer. Statist. Assoc.*, 100:1429–1442.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1):1.

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

Opsomer, J. D., Breidt, F. J., Moisen, G., and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*, 102(478):400–409.

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

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.

Santacatterina, M. and Bottai, M. (2018). Optimal probability weights for inference with constrained precision. *Journal of the American Statistical Association*, 113(523):983–991.

Särndal, C.-E. (1980). On the π-inverse weighting best linear unbiased weighting in probability sampling. *Biometrika*, 67:639–650.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Särndal, C.-E. and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian J. of Statistics*, 11:146–156.

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

Scornet, E. (2016b). Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62:1485–1500.

Scornet, E. (2017). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys*, 60:144–162.

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

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

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC BioInformatics*, 25(8).

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. (2014). Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*.

Wang, L. and Wang, S. (2011). Nonparametric additive modelassisted estimation for survey data. *Journal of Multivariate Analysis*, 102:1126–1140.

Wright, M. and Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185–193.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science*, 3(2):625–650.