Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Design consistent random forest models for data collected from a complex sample

by Daniell Toth and Kelly S. McConville

Release date: December 20, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2024
Vol. 50, No. 2, pp. 185-207
Statistics Canada, Catalogue No. 12-001-X

Design consistent random forest models for data collected from a complex sample

Daniell Toth and Kelly S. McConville[1]

Abstract

Random forest models, which are the result of averaging the estimated values from a large number of tree models, represent a useful and flexible tool for modeling the data nonparametrically to provide accurately predicted values. There are many potential applications for these types of models when dealing with survey data. However, survey data is usually collected using an informative sample design, so it is necessary to have an algorithm for creating random forest models that account for this design during model estimation.

The tree models used in the forest are typically obtained by estimating tree models on bootstrapped samples of the original data. Since the models depend on the observed data and the values observed in the sample depend on the informative sample design, the usual method for estimation is likely to lead to a biased random forest model when applied to survey data.

In this article, we provide an algorithm and a set of conditions that produce consistent random forest models under an informative sample design and compare this method to the usual random forest modeling method. We show that ignoring the design can lead to biased model estimates.

Key Words: Machine learning; Nonparametric; Sample design; Survey data; Tree models.

1. Introduction

Recursive partitioning algorithms were first suggested by Morgan and Sonquist (1963) as a method for analyzing survey data because of the complicated relationships, including interaction effects, among variables that are typical of these datasets. Variables collected from a survey are often highly correlated (even collinear) with each other, are frequently categorical, and can contain many missing values. These complications can make it difficult to make inference about the target population with this data using traditional parametric models. Tree models, which are estimated by applying a recursive partitioning algorithm to the dataset, handle this type of data easily. The variables used in the model along with any interaction effects are selected automatically and the resulting binary split structure makes these models easy to interpret and identifies complicated interaction effects between the variables in the dataset (Phipps and Toth, 2012; Earp, Toth, Phipps and Oslund, 2018).

Given a set of n observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^{n}$, of a random response variable Y and d random predictor variables $\mathbf{X} = (X_1, \ldots, X_d)$, from an informative sample, we want to estimate k new values $\{y_i,\}_{i=n+1}^{n+k}$, given predictor values $\{\mathbf{x}_i\}_{i=n+1}^{n+k}$ for non-sampled units in the population. By estimating the mean function $E[Y \mid \mathbf{X} = \mathbf{x}] = h(\mathbf{x})$ from the observed data we can get predictions of y_i from the model $\tilde{y}_i = \tilde{h}(\mathbf{x})$.

Using survey data to estimate a model in order to obtain good predictions rather than estimates of a finite population parameter is a topic with increasing interest (Wieczorek, 2023). For example, Hong and He (2010) use longitudinal study data to fit a model that can be used to predict the functional mobility status

1. Daniell Toth, Office of Survey Methodology, U.S. Bureau of Labor Statistics. E-mail: toth.daniell@bls.gov; Kelly S. McConville, Department of Statistics, Harvard University.

among the elderly. Meanwhile, Kshirsagar, Wieczorek, Ramanathan and Wells (2017) and Krebs, Reeves and Baggett (2019) both use machine learning models to predict poverty levels and under-story vegetation structure respectively. Similarly, we are interested in estimating the regression function using a machine learning approach using data from an informative sample design. Like Nalenz, Rodemann and Augustin (2024) we propose an approach to modeling random forests using survey data.

A tree model, $\tilde{h}(\mathbf{x})$, is a nonparametric model obtained from an algorithm that recursively partitions the observed data and then estimates the desired statistic for each final partitioning box (end node) separately. The recursive partitioning algorithm consists of choosing a variable X_j among all the available d variables given by vector $\mathbf{X}$ and a value a in which to split the set of observations into two nodes: the observations where $\mathbf{x}_j \le a$ and $\mathbf{x}_j > a$. This procedure is then repeated for each node until there are not enough observations to split or some stopping criteria have been reached (Hothorn, Hornik and Zeileis, 2006). This algorithm results in a set of q boxes, $Q^n = \{B_1, \ldots, B_q\}$ which completely partition the support of $\mathbf{X}$ and depend on the values of the observed data.

Though tree models are easy to interpret, making them ideal for many inference applications, they are not very efficient models for producing point estimates. They are particularly inefficient for models that have linear effects (Loh, 2008). A random forest model, in contrast to the easily interpretable tree model, estimates the expected value of the response variable conditionally on the predictor variables, by averaging the estimates from a set of regression tree models.

Given a set of M regression tree models, $\{\tilde{h}_j\}_{j=1}^{M}$, the random forest estimator of $h(\mathbf{x})$ is

$$\mathcal{F}_0(\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M}\tilde{h}_j(\mathbf{x}), \tag{1.1}$$

where each tree model, $\tilde{h}_j$, is fit using a random subset of the predictor variables on a bootstrap sample of the observed data (Breiman, 2001). Though these models lose the feature of easy interpretation that tree models possess, they are known to provide very accurate predictions and still retain the applicability to a wide range of data types (Breiman, 2001). This provides a useful and flexible tool for accurately modeling the response variable of a given dataset which could have many applications in analyzing data from an informative sample.

For example, Buskirk (2018) and Bilton, Jones, Ganesh and Haslett (2017) present applications of regression trees and random forests on data collected using a complex sample design. Unfortunately, the standard random forest algorithms are meant for independent and identically distributed (i.i.d.) data and many surveys use a complex sample design to collect observations, violating the i.i.d. assumption and in many applications of tree-based models, the available sample design information is often ignored, likely leading to biased estimates as demonstrated by the results in Toth and Eltinge (2011).

Dagdoug, Goga and Haziza (2021) extended the work of McConville and Toth (2019) by using a forest model instead of a single tree as the assisting model in a model assisted estimator to estimate a finite population total. They point out that, if the variables used to determine the sample design are available, it is

possible much of the bias could be reduced by including them in the model. These variables are extremely useful for estimation of population parameters in the context of model assisted estimation, but these variables are not available for making predictions of values for units outside of the sample.

It is desirable then to have an algorithm that allows consistent estimation of a regression function for the population, estimated using data from an informative design, that can be used for prediction. For example work being done at the BLS requires a model to predict respondent-burden for households selected in the Consumer Expenditure Survey from household characteristics which are believed to be associated with burden Yang and Toth (2022).

In this article, we propose a design consistent random forest model for the regression function that uses a weighted average of end-nodes obtained from a set of purely random trees that incorporate sampling weights in their estimation. This process avoids having to produce sensible bootstrap samples from a general sample design. Forests constructed from completely random trees have been studied in the literature with the method commonly referred to as the *uniform random forest algorithm* (Biau, Devroye and Lugosi, 2008; Scornet, 2016; Arlot and Genuer, 2014) but these models are generally not effective in practice. They are used primarily to study the theoretical properties and to understand the behavior and limits of ensemble methods. In the standard uniform random forest algorithm, the end-node estimates are simple averages whereas the end-node weights of our method enables good predictive properties.

To our knowledge our algorithm is the first to propose using weights at the end-node, rather than at the tree level. Because the trees are produced using completely random splits of the predictor space all the real work comes from these weights, providing an adaptive and more efficient estimator. We show that this model provides design consistent estimates and is, therefore, more appropriate for use with data collected using an informative sample design.

In Section 2 we introduce the tree model and provide the necessary assumptions for its design consistency. Section 3 contains the method for using tree models in a random forest model that requires weighting each tree model estimate and the statement of the main theoretical result of the article, which is that the proposed random forest estimator is design consistent estimator of the regression function. Appendix contains all necessary auxiliary lemmas and proofs of the results. Section 4 summarizes simulation studies where we compare the performance of our proposed method to the standard random forest estimator on data from simple random samples (SRS) and from probability proportional to size (PPS) samples. In particular, we apply our proposed random forest model to the Academic Performance Index (API) score data from standardized test results of students computed for all California schools with at least 100 students and the U.S. Bureau of Labor Statistics' (BLS) Consumer Expenditure (CE) data. These results demonstrate that ignoring the sample weights using data from an informative sample design could lead to biased forest estimators. Lastly, also in Section 4, we conduct a simulation study on generated data to explore consistency of our proposed method and the standard random forest model.

2. Design consistent tree models

Consider a finite population of size N, $\{(Y_i, \mathbf{X}_i, \mathbf{Z}_i)\}_{i=1}^{N}$, generated from a super-population model ξ, where Y_i is the variable of interest associated with unit i and $\mathbf{X}_i$ is a d -vector of potential predictors variables associated with unit i that are part of the released data available to the analyst. We use $\mathbf{Z}_i$ to denote a d^*-vector of variables associated with unit i, known by the survey designer but not released with the survey data for analysis.

A random sample $S \subset U = \{1,\ldots,N\}$ of size n is selected using a sample design with inclusion probabilities $\pi_i = P(i \in S)$. The sample design, defined by the inclusion probabilities, can depend on variables associated with the unit where some are known and some are unknown to the analyst, $\pi_i = P(i \in S \mid \mathbf{x}_i, \mathbf{z}_i)$. If these inclusion probabilities are associated with Y, the design can affect the estimates and inference for the population that result from using the sample data. Such sample designs are called *informative* sample designs. If the sample design only depends on variables available to the data analyst, $\pi_i = P(i \in S \mid \mathbf{x}_i, \mathbf{z}_i) = P(i \in S \mid \mathbf{x}_i)$, design-consistent models may be able to be obtained by incorporating all of these variables used to define the inclusion probabilities into the modeling process (Gelman, King and Liu, 1998; Little, 2004). However, in most publicly released survey datasets, many of the variables used in the survey design are not released to the data analyst. Instead, the data is released with a set of survey weights $\{w_i\}_{i \in S}$ intended to be used by the data analyst to account for the survey design in the analysis (Lavallée and Beaumont, 2015; Pfeffermann, 1993). Besides accounting for the probability of selection into the sample, these weights often include adjustments for nonresponse and/or known totals of key auxiliary information. Though our arguments in this article could be used for general survey weights, for simplicity of exposition, we assume that the weight associated with unit i is the inverse of the probability of selection for that unit, π_i^{-1}.

In order to study large sample properties of the estimator in this context, it is necessary to consider a sequence of populations that are increasing in size and distributed i.i.d. from super-population and a sequence of associated sample designs. From each population-design pair a corresponding sequence of samples is drawn, also increasing in size and each selected according to the sample design. More concretely, suppose we have a sequence of finite populations $\{(Y_1, \mathbf{X}_1), \ldots, (Y_{N_\nu}, \mathbf{X}_{N_\nu})\}$, indexed by ν, so that $U_1 \leq \ldots, \leq U_\nu$, with sizes $N_1 \leq \ldots \leq N_\nu$. Each finite population is generated by taking i.i.d. draws from the distribution of the super-population ξ. The random samples, $S_1 \subset U_1, \ldots S_\nu \subset U_\nu$, are drawn from each finite population using the corresponding sample design, with increasing sizes $n_1 \leq \ldots \leq n_\nu$. It is the behavior of the sequence of estimates obtained from these samples that is considered.

If a tree model, $\tilde{h}_\nu(\mathbf{x})$, is obtained by recursively partitioning the observed sample data and then estimating the mean in each box, the resulting tree model is an estimator of the conditional mean function $h(\mathbf{x}) = E_\xi[Y \mid \mathbf{x}]$. Toth and Eltinge (2011) provide an algorithm for estimating $h(\mathbf{x})$ and a set of conditions for which this estimator is an L^2 consistent estimator of $h(\mathbf{x})$. In this article, we intend to propose a random forest model that is constructed from a weighted average of these design consistent tree models. For the rest of this section, we review the notation and results necessary to establish a design consistent algorithm for

random forest models including a discussion of the conditions and the main result for the tree model given in Toth and Eltinge (2011).

Let $Q^{n_\nu} = \{B_1^{n_\nu}, \ldots, B_q^{n_\nu}\}$ be the set of partitioning boxes that result from applying a recursive partitioning algorithm to the observed sample S_ν. To facilitate discussion of the predicted value of an observation with predictor variables $\mathbf{x}$ for a given tree, we now define some functions that help simplify the notation. Let $B^{n_\nu}(\mathbf{x})$ denote the box in Q^{n_ν} containing the value $\mathbf{x}$. The functions $\#B^{n_\nu}(\mathbf{x}) = \sum_{i \in S} \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}$ and $\tilde{\#}B^{n_\nu}(\mathbf{x}) = \sum_{i \in S} \pi_i^{-1} \mathbb{I}_{\{x_i \in B^{n_\nu}(\mathbf{x})\}}$ provide the number of observed sample units and estimated number of population units in box $B^{n_\nu}(\mathbf{x})$ respectively. The estimated mean of the box containing the value $\mathbf{x}$ is define as

$$\tilde{\mu}(\mathbf{x}) = \left[\tilde{\#}B^{n_\nu}(\mathbf{x})\right]^{-1} \sum_{i \in S} \pi_i^{-1} y_i \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}. \tag{2.1}$$

Note that this is the standard Hájek estimator of the mean for population units contained in the partitioning box $B^{n_\nu}(\mathbf{x})$ (Hájek, 1960).

Each box B^{n_ν} in a given partition Q^{n_ν} has two corresponding index vectors which define the borders of the box in all d dimensions. Given a box of a partition, B^{n_ν}, let $\mathbf{a}(B^{n_\nu}) = (a_1(B^{n_\nu}), \ldots, a_d(B^{n_\nu}))$ and $\mathbf{b}(B^{n_\nu}) = (b_1(B^{n_\nu}), \ldots, b_d(B^{n_\nu}))$, where for every $\mathbf{x} \in B^{n_\nu}$, $a_l(B^{n_\nu}) \le x_l < b_l(B^{n_\nu})$, for $l = 1 \ldots d$.

For a given value $\mathbf{x}$ in the support of $\mathbf{X}$, we use the notation $\tilde{F}_i^{\nu}(\cdot)$ to denote the empirical marginal distribution function of x_l conditioned on the partition. That is, for a constant c, and given value $\mathbf{x}$ the empirical marginal distribution function of x_l conditioned on the partition is

$$\begin{aligned} \tilde{F}_l^{\nu}(c \mid Q^{n_\nu}) &= \tilde{F}_l^{\nu}\left(c \mid B^{n_\nu}(\mathbf{x})\right) \\ &= \left(\tilde{\#}_{N_\nu}\left(B^{n_\nu}(\mathbf{x})\right)\right)^{-1} \sum_{i \in S_\nu} \pi_{\nu i}^{-1} \mathbb{I}_{\{x_{li} \le c\}} \mathbb{I}_{\{\mathbf{x}_i \in B^{n_\nu}(\mathbf{x})\}}. \end{aligned} \tag{2.2}$$

The left continuous conditional empirical marginal distribution function $\tilde{F}_l^{-}$ is defined by replacing the indicator function $\mathbb{I}_{\{x_{li} \le x\}}$ in the above definition with $\mathbb{I}_{\{x_{li} < x\}}$. We will also use the empirical probability function $\tilde{P}_{n_\nu}(\mathcal{A})$ of a given event $\mathcal{A}$. The empirical probability function is defined as

$$\tilde{P}_{n_\nu}(\mathcal{A}) = \tilde{N}_\nu^{-1} \sum_{i \in S_\nu} \pi_{\nu i}^{-1} \mathbb{I}_{\{\mathcal{A}\}}(\mathbf{x}_i), \tag{2.3}$$

where $\mathbb{I}_{\{\mathcal{A}\}}(\mathbf{x}_i) = 1$, if the event $\mathcal{A}$ is satisfied for observation $\mathbf{x}_i$ and where $\tilde{N}_\nu = \sum_{i \in S_\nu} \pi_i^{-1}$.

Next we define the l -norm of partition Q^{n_ν} relative to $\tilde{F}_l$ by

$$\| Q^{n_\nu} \|_l^{\tilde{F}} = \sum_{B^{n_\nu} \in Q^{n_\nu}} \left\{ \left[\tilde{F}_l\left(b_l(B^{n_\nu})\right) - \tilde{F}_l\left(a_l(B^{n_\nu})\right) \right] \tilde{P}(\mathbf{x} \in B^{n_\nu}) \right\} \tag{2.4}$$

and the l -norm of partition Q^{n_ν} relative to $\tilde{F}^{-}$

$$\| Q^{n_\nu} \|_l^{\tilde{F}^{-}} = \sum_{B^{n_\nu} \in Q^{n_\nu}} \left\{ \left[\tilde{F}_l^{-}\left(b_l(B^{n_\nu})\right) - \tilde{F}_l^{-}\left(a_l(B^{n_\nu})\right) \right] \tilde{P}(\mathbf{x} \in B^{n_\nu}) \right\}. \tag{2.5}$$

The following conditions on the super-population model, sample design and partition created from the algorithm are sufficient to show that a regression tree model based on the sample data is an L^2 - consistent estimator of the true conditional mean of the variable of interest, Y. In Section 3, we show that these conditions (with the strengthening of one condition) are sufficient to obtain consistent forest estimators as well. The proofs for consistent regression trees require only a finite second moment of the variable of interest, but we require a finite fourth moment to prove consistency of the proposed forest estimator.

Many of the conditions to obtain consistency require understanding the rate at which things converge. Before specifying the conditions on the population, sample design and algorithm, we first define two scalar functions that will be used as the rates of convergence. Let $\gamma(x)$ and $k(x)$ be functions bounded above 0 for all $x > 0$ satisfying:

1: $\gamma(x) \to \infty$

2: $x^{-1}k(x) \to 0$

3: $k(x)^{-1}\gamma(x)x^{1/2} \to 0,$

as $x \to \infty$. These constraints require both functions to be unbounded where $\gamma(x)$ grows to ∞ slower than $\sqrt{x}$, while $k(x)$ grows faster than $\sqrt{x}$, but slower than x. Note that there are an infinite number of function pairs that satisfy these three constraints. Below we use these functions to specify the relative speeds at which different terms converge relative to the sizes of the population N_v and sample n_v. We will also use the sampling fraction, defined as $f_v = n_v / N_v$.

Condition 1: $\lim_{v\to\infty} N_v^{-1}\sum_{i=1}^{N_v} Y_i^4 < \infty$

Condition 2: $\limsup_{v\to\infty}(N_v \min_{i\in U_v}\pi_{vi})^{-1} = O(n_v^{-1})$

Condition 3: $\limsup_{v\to\infty}\max_{i,j\in U_v, i\neq j}\left|\frac{\pi_{vij}}{\pi_{vi}\pi_{vj}} - 1\right| = O(N_v^{-1})$

Condition 4: $f_v^{-1} = O(n_v^{1/2}\gamma(n_v)^{-1})$

Condition 5: $E_p[\delta_{vi}\delta_{vj} \mid Q^{n_v}] = \pi_{vij} \quad \forall i, j \in U_v$

Condition 6: $\tilde{P}\left(k(n_v)^{-1}\#_{n_v}(B^{n_v}(\mathbf{x})) \geq 1\right) \to_p 1$

Condition 7: $\| Q^{n_v} \|_l^{\tilde{F}_{n_v}} \to_p 0$ and $\| Q^{n_v} \|_l^{\tilde{F}_{n_v}^-} \to_p 0,$ for $l = 1, \ldots, d$

where the above conditions are all assumed with ξ -probability 1.

Condition 1 is the only condition directly on the distribution of the super-population model. The data do not need to follow any predefined distribution, requiring only that the outcome variable, Y, has a finite fourth moment. This generality makes these models applicable to a wide class of problems. We use this condition on the fourth moment for establishing design consistency of the proposed forest estimator, but as mentioned above, this condition could be weakened to require only a finite second moment in the case of design consistent tree estimators.

Conditions 2 through 4 are standard conditions on the sample design requiring that every unit in the population can be selected with some minimum probability, the effect of clustering shrinks relative to the population size and a mild requirement on the sampling rate (Isaki and Fuller, 1982; Breidt and Opsomer, 2000). Condition 4 is a weak limit on how big the finite populations can grow relative to the sample size in the sense that it allows for an arbitrarily small sampling rate.

Condition 5 is a condition from Toth and Eltinge (2011) requiring that the selection probabilities are independent of the given partition. The partitioning set Q^{n_ν} is a function of an algorithm applied to the selected data, so this condition on both the algorithm and the sample design limits the influence any selected unit can have on the resulting partition.

Condition 6 and 7 are both conditions on the partitioning algorithm. The first requires that the number of observations in each partitioning box grows at a certain rate relative to the sample size, while the second requires the l-norms of the partitioning boxes, defined by 5 and 6, shrink toward zero as the sample size grows.

Proposition 2.1 (Toth and Eltinge, 2011)**.** *Let* $\{(Y_i, \mathbf{X}_i)\}_{i\in U_\nu}$, *be a sequence of finite populations, indexed by* ν *and distributed i.i.d. from the super-population model* ξ *and let* S_ν *denote a random sample from* U_ν *selected using the sample design. Given* Q^{n_ν}, *the collection of partitioning boxes created from the algorithm applied to the sample data,* S_ν, *define*

$$\tilde{h}_{n_\nu}(\mathbf{x}) = (\tilde{\#}B^{n_\nu}(\mathbf{x}))^{-1}\sum_{i\in S_\nu}\pi_i^{-1}y_i\,\mathbb{I}_{\{\mathbf{x}_i\in B^{n_\nu}(\mathbf{x})\}}. \tag{2.6}$$

If $\lim_{\nu\to\infty}N_\nu^{-1}\sum_{i=1}^{N_\nu}Y_i^2<\infty$ *and Conditions 2 through 7 are satisfied with* ξ *-probability 1, then*

$$\lim_{\nu\to\infty}E_{\xi p}\left[\left|\tilde{h}_{n_\nu}(\mathbf{x})-h(\mathbf{x})\right|^2\right]=0.$$

Notice that the right hand side of equation (2.6) is the Hájek estimator of the mean of the box containing $\mathbf{x}$ given by (2.1). Since the set of boxes partition the data, each observation falls into exactly one box and the model predicted value of Y for an observation with auxiliary variables $\mathbf{X}=\mathbf{x}$ is simply

$$\tilde{h}_{n_\nu}(\mathbf{x}) = \tilde{\mu}(\mathbf{x}). \tag{2.7}$$

Therefore, Proposition 2.1 tells us that a regression tree model that estimates the mean of Y in each end node is an L^2 - consistent estimator of the function $E_\xi[Y\,|\,\mathbf{x}]$, so

$$\tilde{\mu}(\mathbf{x})\to_{L^2}E_\xi[Y\,|\,\mathbf{x}]. \tag{2.8}$$

We will rely on this result in the following section to show that the proposed random forest estimator is consistent as well as the following corollary.

Corollary 2.1. *For a given tree* j, *if Conditions 1 through 7 are satisfied, then*

$$\left[\tilde{\#}B_j^{n_v}(\mathbf{x})\right]^{-1}\sum_{i\in S}\pi_i^{-1}y_i^2\,\mathbb{I}_{\{\mathbf{x}_i\in B_j^{n_v}(\mathbf{x})\}}\to_p E_\xi[Y^2\mid\mathbf{x}].$$

Proof in Appendix.

3. Design consistent forest models

Random forest model estimates are obtained by averaging the estimates of M tree models. This requires using a procedure for producing several *different* tree models using the same data; it does not improve the estimate to average over the same model. The tree models used in the forest are typically obtained by estimating tree models on bootstrapped samples of the original data and using a random subset of the predictor variables at each split (Breiman, 2001). However, a bootstrap sample of a dataset is not always easy or possible to produce for a general informative sample design (Mashreghi, Haziza and Léger, 2016). The typical approach of ignoring the sample design during estimation is likely to lead to a biased random forest model when applied to survey data.

That is, given an estimator $\hat{m}_n$ of the regression function m and a point $\mathbf{x}$,

$$\text{Bias}(\hat{m}_n(\mathbf{x})) := \mathbb{E}[\hat{m}_n(\mathbf{x})]-m(\mathbf{x}), \tag{3.1}$$

where the expectation is taken with respect to the joint distribution of the super-population model and the sample design.

We now propose a forest model that is design consistent for a family of informative sample designs provided that the sample design, super-population distribution, and recursive partitioning procedure satisfy Conditions 1 through 7 in Section 2. In order to obtain different regression tree models from a given sample, at each step of the recursive partitioning we select the variable completely at random from the d possible predictor variables and the splitting point at random from the observed support of the selected variable. This algorithm is outlined in Figure 3.1.

Figure 3.1 Recursive partitioning algorithm to produce random tree models.

Recursive Partitioning Algorithm
1. Let $n_{\text{end}} = \max(5,\ \text{floor}\ \{10^{-7}n\})$.
2. If the dataset contains at least $2n_{\text{end}}$ observations continue to the next step; otherwise stop.
3. Among the auxiliary variables x_l, $l=1,\ldots,d_1$, randomly choose a variable on which to split the data.
4. Split the data into two sets S_L and S_R by randomly selecting a value of the selected variable x_l that results in each sub-dataset containing at least n_{end} observations.
5. Apply the algorithm beginning at step 2 to each of the two resulting subsets S_L and S_R.

Note: Notice that n_{end} is defined so that the end-nodes of each tree satisfies Conditions 6 and 7.

Note that n_{end} is defined in such a way as to satisfy Condition 6, because it is linear in n and thus dominates $\sqrt{n}$, but still allows for a relatively small number of observations. This is important because in practice a small number of observations is effective in getting accurate estimates.

3.1 Extended notation for forests

Because we are interested in forests, which require a set of trees, we extend some of the notation and functions used in Section 2 to facilitate this discussion. For instance, rather than one set of partitioning boxes for instance v, we will have one for each of the M trees in the model. Denote the partition for the j-th tree by $Q_j^{n_v} = \{B_{j1}^{n_v}, \ldots, B_{jq_j}^{n_v}\}$. Note that the number of partitioning boxes q, created using the sample S_v, can be different for the different trees in the forest model, so depends on j. Let $\mathcal{Q}^{n_v} = \{Q_j^{n_v}\}_{j=1}^{M}$ be the set of all the partitions making up the forest model.

The function $B_j^{n_v}(\mathbf{x})$ will denote the box in the j-tree that contains the value $\mathbf{x}$, while $\#B_j^{n_v}(\mathbf{x})$ and $\tilde{\#}B_j^{n_v}(\mathbf{x})$ are the number of observed sample units and estimated number of population units in box $B_j^{n_v}(\mathbf{x})$ respectively. Likewise, the estimated mean of the observations in the box containing the value $\mathbf{x}$, defined by equation (2.1), for the j-th tree is denoted $\tilde{\mu}_j(\mathbf{x})$.

3.2 Weights for averaging estimates

Notice that the algorithm and therefore the structure of each tree depends only on the observed values of the modeling variables $\{\mathbf{X}_i\}_{i=1}^{n}$, resulting in a k-nearest neighbor estimate of Y based on the closeness of a random sub-sample of modeling variables. The forest model is then an average over M k-nearest neighbor estimates. However, because the trees are built based on randomly selected splits, the homogeneity of Y will likely vary across boxes, resulting in more or less informative boxes. So while the simple average of the random tree estimates given by (1.1) leads to an asymptotically unbiased estimator, it will also be rather inefficient.

In order to improve the efficiency of the forest estimator, we use a weighted average, with the goal of giving more weight to estimates from tree models with greater predictive accuracy. Let $\{\tilde{h}_j^{n_v}\}_{j=1}^{M}$, denote the set of regression tree models. Then a weighted forest model would have the form

$$\mathcal{F}_w(\mathbf{x}) = \sum_{j=1}^{M} \lambda_j(\mathbf{x})\, \tilde{h}_j^{n_v}(\mathbf{x}), \tag{3.2}$$

where $\lambda_j(\mathbf{x})$ is a weight that depends on the end-node of tree j that $\mathbf{x}$ belongs to.

Methods for using a weighted average of tree estimates to produce a forest estimate have been considered (Gajowniczek, Grzegorczyk, Ząbkowski and Bajaj, 2020; Shahhosseini and Hu, 2020; Winham, Freimuth and Biernacka, 2013) but these involve using a weight based on the fit of each tree only. In testing several different approaches, we found using a weight that depends on the final end node produced the best results. However, this approach induces bias in the estimates which needs to be adjusted for.

For our proposed method, we weight each tree estimate using a weight that is inversely proportional to the estimate of one plus the end node variance, $V_{B_j^{n_v}(\mathbf{x})} = \mathrm{Var}_{\xi}(Y \mid \mathbf{X} \in B_j^{n_v}(\mathbf{x}))$ similar to resulting weights used in some adaptive methods (Williams, Neilley, Koval and McDonald, 2016). If we knew the true mean, $\mu_j(\mathbf{x}) = E[Y \mid B_j^{n_v}(\mathbf{x})]$, of the Y-values for the observations in box $B_j^{n_v}(\mathbf{x})$, then a design consistent estimator of $V_{B_j^{n_v}(\mathbf{x})}$ is given by

$$\left[\tilde{\#}B_j^{n_v}(\mathbf{x})\right]^{-1} \sum_{i \in S} \pi_i^{-1} (y_i - \mu_j(\mathbf{x}))^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}}.$$

However, because the true $\mu_j(\mathbf{x})$ is unknown, we use the estimator

$$\tilde{V}_{B_j^{n_v}(\mathbf{x})} = \left[\tilde{\#}B_j^{n_v}(\mathbf{x})\right]^{-1} \sum_{i \in S} \pi_i^{-1} (y_i - \tilde{\mu}_j(\mathbf{x}))^2 \, \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_v}(\mathbf{x})\}}, \tag{3.3}$$

where the estimated value $\tilde{\mu}_j(\mathbf{x})$ replaces the true mean.

Given $\mathbf{x} \in B_j^{n_v}(\mathbf{x})$, then the weight for tree j in the forest is set to

$$\lambda_j(\mathbf{x}) = \frac{(\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}}{\sum_{j=1}^{M} (\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}}. \tag{3.4}$$

so that the weights $\lambda_j(\mathbf{x}) \propto (\tilde{V}_{B_j^{n_v}(\mathbf{x})} + 1)^{-1}$ and they sum to 1.

The forest estimated value of y given the $\mathbf{x}$ is

$$\sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{h}_j^{n_v}(\mathbf{x}) = \sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{\mu}_j(\mathbf{x}). \tag{3.5}$$

noting again the equivalence between the j-th tree estimate and the estimated mean of the end-node that contains $\mathbf{x}$ for tree j. For a given set of sample data, each $\lambda_j(\mathbf{x})$ and $\tilde{\mu}_j(\mathbf{x})$ are functions of the random partitioning process, so can be seen as M independent observations of the random vector, $(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x}))$. Therefore, the expression given by (3.5) can be seen as a sum of products of the components of these M random vectors.

3.3 Estimate of bias from weights

Using this weighted average does increase the efficiency of the estimator, but also makes the estimator potentially biased under the sample design. In particular, if we explore the expectation of the weighted forest model with respect to the selection probability and the randomness of the recursive partitioning algorithm (Figure 3.1) denoted by E_{p^*}, then we get

$$\begin{aligned} E_{p^*}\left[\sum_{j=1}^{M} \lambda_j(\mathbf{x}) \, \tilde{h}_j(\mathbf{x})\right] &= \sum_{j=1}^{M} E_{p^*}\left[\lambda_j(\mathbf{x})\right] E_{p^*}\left[\tilde{h}_j(\mathbf{x})\right] + \sum_{j=1}^{M} \mathrm{cov}_{p^*}\left(\lambda_j(\mathbf{x}), \tilde{h}_j(\mathbf{x})\right) \\ &= \tilde{h}^* E_p\left[\sum_{j=1}^{M} \lambda_j(\mathbf{x})\right] + \sum_{j=1}^{M} \mathrm{cov}_{p^*}\left(\lambda_j(\mathbf{x}), \tilde{\mu}_j(\mathbf{x})\right), \end{aligned}$$

where $E_{p^*}[\tilde{h}_j] = \tilde{h}^*$ for each j.

Therefore,

$$E_{p^*}\left[\sum_{j=1}^{M}\lambda_j(\mathbf{x})\,\tilde{h}_j(\mathbf{x})\right] = E_\xi[Y \mid \mathbf{x}] + \sum_{j=1}^{M}\mathrm{cov}_{p^*}(\lambda_j(\mathbf{x}),\,\tilde{\mu}_j(\mathbf{x})),$$

because $\tilde{h}^*(\mathbf{x}) \to E_\xi[Y \mid \mathbf{X} = \mathbf{x}] = E_\xi[Y \mid \mathbf{x}]$ by Proposition 2.1 and $\sum_{j=1}^{M}\lambda_j(\mathbf{x}) = 1$ by design.

Since each $\lambda_j(\mathbf{x})$ and $\tilde{\mu}_j(\mathbf{x})$ are observations of random variables $\lambda(\mathbf{x})$ and $\tilde{\mu}(x)$, for a given value $\mathbf{x}$, the bias term is

$$\sum_{j=1}^{M}\mathrm{cov}(\lambda_j(\mathbf{x}),\,\tilde{\mu}_j(\mathbf{x})) = M\mathrm{cov}(\lambda(\mathbf{x}),\,\tilde{\mu}(\mathbf{x})). \tag{3.6}$$

In order to correct for this bias for a fixed sample, we estimate $\mathrm{cov}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x}))$ using the M observations by

$$\widehat{\mathrm{cov}}(\lambda(\mathbf{x}),\,\tilde{\mu}(\mathbf{x})) \;=\; (M-1)^{-1}\sum_{j=1}^{M}(\lambda_j(\mathbf{x}) - \bar{\lambda})\,(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}), \tag{3.7}$$

where $\bar{\lambda} = M^{-1}\sum_{j=1}^{M}\lambda_j(\mathbf{x})$ and $\bar{\mu} = M^{-1}\sum_{j=1}^{M}\tilde{\mu}_j(\mathbf{x})$. Therefore the proposed forest estimator for the function $h(\mathbf{x})$ is

$$\mathcal{F}_{n_v}(\mathbf{x}) = \sum_{j=1}^{M}\lambda_j(\mathbf{x})\,\tilde{\mu}_j(\mathbf{x}) - M\widehat{\mathrm{cov}}(\lambda(\mathbf{x}),\,\tilde{\mu}(\mathbf{x})). \tag{3.8}$$

The following result is the main theoretical result of the article. It states that the proposed forest estimator, $\mathcal{F}_{n_v}(\mathbf{x})$, defined by equation (3.8) is asymptotically design-unbiased which converges in probability to $h(\mathbf{x}) = E_\xi[Y \mid \mathbf{X} = \mathbf{x}]$. This result provides theoretical justification for using this random-forest estimation method on data collected from an informative sample design.

Proposition 3.1. *For a fixed $M > 0$, if Conditions 1 through 7 are satisfied for each tree in the forest, then*

$$\mathcal{F}_{n_v}(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}],$$

for all $\mathbf{x}$ as $n_v \to \infty$.

4. Relative performance of the estimators

In order to understand how the proposed random forest method compares to the typical i.i.d. random forest of Breiman (2001), we test the two methods over repeated samples of two publicly available datasets using two different sample designs, simple random sampling (SRS) and sampling with probability proportional to size (PPS). We evaluate the efficiency and bias of the predictions of each methods empirically and compare them to the standard Hájek estimator. The proposed random forest approach given by the algorithm in Figure 3.1, was tested using the algorithm available in the R-package *rpms* (Toth, 2024),

and for the method (RF), proposed by Breiman (2001), we use the algorithm available in the R-package *randomForest* (Liaw and Wiener, 2002).

For the finite populations, we use the two datasets described below. In each dataset description, we also identify the variable of interest, the predictor variables, and the variable used as the measure of size for the PPS sample design.

API The California Academic Performance Index (API) dataset available in the *survey* package (Lumley, 2020) contains data on 5,973 schools including the school average score on the API standardized test as well as demographic and administrative data about the school and the neighborhood it serves. We treat the school's average test score in the 2000 academic year as the variable of interest with five school level predictor variables including 1.) the average education level attained by the parents of the students in the school, 2.) the percentage of students that are English language learners, 3.) the percentage of students enrolled in a subsidized lunch program, 4.) the percentage of teachers with full qualifications, and 5.) whether or not the school was eligible for an awards program. To compare the methods on an informative PPS design, we sample proportionally by the size of the school's enrollment.

CEx A subset of the Consumer Expenditure Survey public use interview data file from the US Bureau Labor of Statistics which is available in the *rpms* package. This dataset contains information from 2015 on 45,308 households with total expenditures greater than \$0. We consider the total household expenditures for the current quarter as the variable of interest with five predictor variables including 1.) whether or not the household lives in a home which they own (with or with a mortgage), rent, or is part of student housing; 2.) the region in which the household is located; 3.) whether or not the household lives in an urban location; 4.) whether or not a member of the household currently earns a wage; and 5.) the age of the person identified as the primary earner of the household. To compare the methods using an informative PPS design, we sample proportionally to the size (number of residents) of the household.

Treating each dataset as a finite population, we take $D = 500$ repeated samples of size n from the population where $n = 600$ for API and $n = 1{,}000$ for CEx. For each random sample, we fit the random forest models with 500 trees to the sample data using the default settings of both algorithms and all the available predictor variables. The default settings require that each end node of every tree contains at least 5 observation. Using these models we predict the values of the variable of interest for every unit in the finite population and the predicted values are compared to the true values.

Specifically, for each sample s_l, $l = 1, \ldots, D$, we find the estimated model $\tilde{h}^{(l)}(\mathbf{x})$ and calculated the empirical mean error

$$b_l = \frac{1}{N} \sum_{i=1}^{N} (\tilde{h}^{(l)}(\mathbf{x}_i) - y_i), \tag{4.1}$$

the empirical mean relative error, $b_l / \bar{y}$, where $\bar{y} = N^{-1}\sum_{i=1}^{N} y_i$, and the empirical mean squared error

$$m_l = \frac{1}{N}\sum_{i=1}^{N}(\tilde{h}^{(l)}(\mathbf{x}_i) - y_i)^2. \tag{4.2}$$

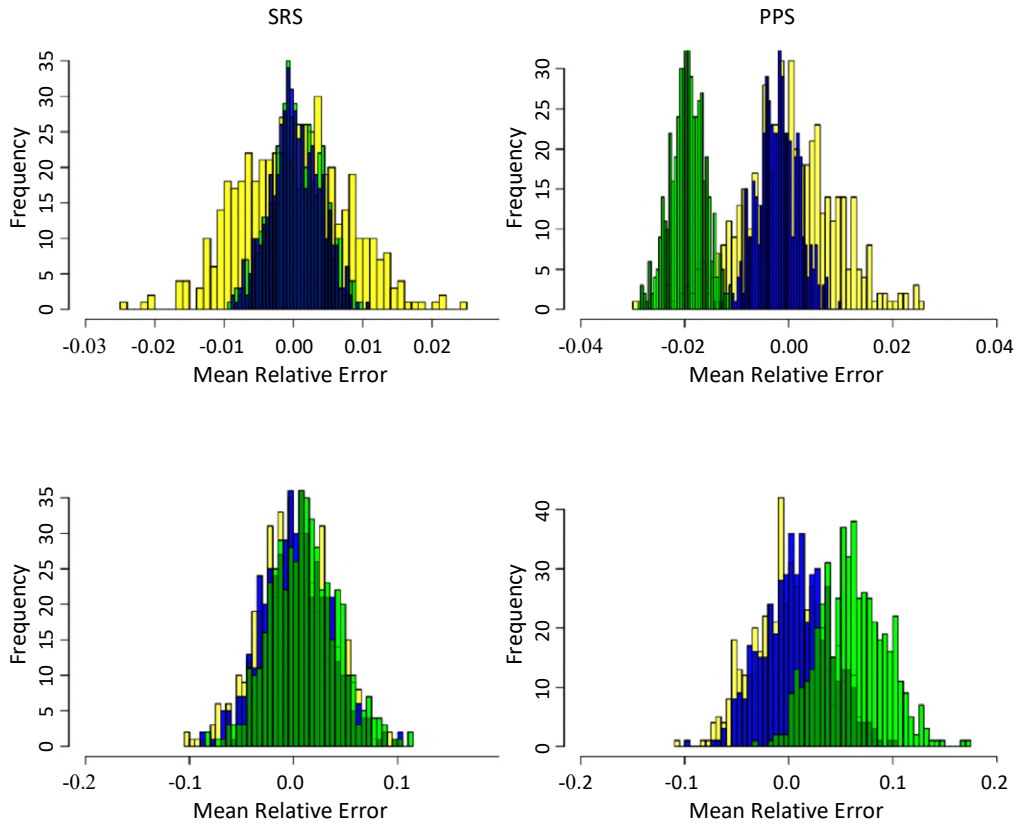
Notice the empirical mean error is an estimate of the average bias, which is defined as

$$\text{ABias}(\hat{m}_n(\mathbf{x})) := \mathbb{E}_{\xi}[\mathbb{E}_{\xi_p}[\hat{m}_n(\mathbf{x})] - m(\mathbf{x})], \tag{4.3}$$

where the first expectation is taken with respect to the joint distribution of the population and sample distribution and the second is with respect to the population distribution.

Since on of the biggest risk of ignoring an informative sample design when modeling a dataset is the introduction of bias in the model, it important to assess the potential bias of each of the estimators. The empirical distributions of the empirical relative mean errors over repeated samples for two data sets using two sample designs for the two forest algorithms as well as the Hájek estimator are shown in Figure 4.1.

Figure 4.1 Distribution of the mean relative errors of the three estimators over 500 repeated samples from the two datasets.



Note: The yellow histogram is the distribution of the RME for the Hájek estimator, the green is the unweighted random forest and the blue is the weighted forest method. The top graphs are the distributions of the mean relative errors using the API dataset and the bottom two using the CEx dataset.

When an SRS design is used, the two distribution of relative error given on the left side of Figure 4.1, show that all three estimators produce relatively unbiased estimates as their errors are centered very close to zero. In addition, the distribution of the errors of the two forest models have a smaller range than the Hájek estimator which shows that using these models leads to an increase in efficiency. This gain of efficiency is especially seen in the distributions using the API data.

Since the RF algorithm ignores the design weights, one would expect that there could be more bias from estimates of values using this model compared to values obtained from the Hájek estimator or the forest model using the algorithm in the *rpms* package, when the sample design is informative. The plotted distributions of the mean relative errors over repeated PPS samples, shown on the right hand side of Figure 4.1, confirm this. Under repeated PPS samples, the Hájek estimator still appears unbiased and the distribution of the relative errors are still wider than both forest models. Though the distribution of the relative errors of the RPMS model (in blue) appears to be centered close to zero, the relative errors of the RF model ignoring the weights is centered close to -2% for the API dataset and around 6% for the CEx dataset. This suggests that ignoring the weights leads to much more bias than the proposed method under the PPS design.

Table 4.1 contains the average relative mean errors and average mean squared error $\bar{m} = D^{-1}\sum_{l=1}^{D} m_l$ over the 500 random samples for each of the three models, the two datasets, and the two sample designs. The relative mean error statistic is presented as a percentage, $(\bar{b} / \bar{y})$ 100% and the mean square error statistic is given relative to that of the Hájek estimator, $\bar{m} / \bar{m}_H$, where $\bar{m}_H$ is the average mean squared errors of the Hájek estimator over the 500 samples.

Table 4.1
Averages over 500 random samples comparing the prediction error using the Hájek estimator and the two random forest methods on two datasets and for two sample designs.

	API N = 5,973 n = 600				CEx N = 45,308 n = 1,000			
	% Rel. Error		RMSE		% Rel. Error		RMSE	
Method	SRS	PPS	SRS	PPS	SRS	PPS	SRS	PPS
Hájek	-0.010	-0.007	1.000	1.000	-0.206	0.066	1.000	1.000
RF	0.043	-1.945	0.209	0.218	0.806	6.111	0.862	0.865
RPMS	0.021	-0.210	0.204	0.204	-0.056	0.878	0.844	0.844

Note: The percent relative error is the mean error of the estimated values for the full population relative to the population mean of the variable of interest, multiplied by 100. The relative RMSE is the mean over the 500 samples of the calculated mean squared error of the estimated values for the full population, relative to that of the Hájek estimator.
API = Academic Performance Index; CEx = Consumer Expenditure Survey; PPS = Probability proportional to size; RPMS = Recursive Partitioning for Modeling Survey Data; SRS = Simple random samples.

One can see from the results in Table 4.1 that relative mean squared error of the two estimators using the forest modeling methods are smaller than that of the Hájek estimator for both datasets under the SRS and PPS sample designs. However, the RF procedure that ignores the sample weights produces biased estimates under the PPS design for both datasets while both the proposed random forest modeling procedure and the Hájek estimator provide relatively unbiased estimates.

4.1 Demonstration of consistency

So far, we have been examining the efficiency and bias of our forest estimator compared to the usual i.i.d. random forest algorithm and the standard Hájek mean estimator by focusing on the difference between the model estimates and the true value of the variable of interest y using two real datasets as our finite population. An estimator $\tilde{h}(\mathbf{x})$ is consistent if

$$E_{\xi p}[(\tilde{h}(\mathbf{x}) - E_{\xi}[Y \mid \mathbf{x}])^2] \to 0 \text{ as } \nu \to \infty, \tag{4.4}$$

which requires knowing the true mean function $h(\mathbf{x}) = E_{\xi}[Y \mid \mathbf{x}]$ and letting the size of the sample and population go to ∞. When using a real dataset as the finite population, you don't know $h(\mathbf{x})$. Therefore, to explore consistency we use simulated data $(Y, \mathbf{X})_{i=1}^{N}$, where the values are obtained through random draws from a known distribution and we study the behavior of the estimators for a sequence of sample sizes.

For each randomly generated observation i we generate a random vector

$$\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, V_{i1}, V_{i2}, V_{i3})$$

of 6 independent random variables. Variables X_{i1} and X_{i2} both follow a uniform distribution $U(-10, 20)$, and $X_{i3} \sim U(-100, 200)$ while V_{i1} through V_{i3} are categorical random variables with equal probability among categories. Both V_{i1} and V_{i2} take one of 5 categories and V_{i3} takes one of 14 categories. These are the auxiliary variables available to the analyst for every unit in the population and can be used in the model for the variable of interest Y.

Because random forests are known to be very flexible, nonparametric models, rather than testing this method on a set of standard parametric models, we show the results for data that follows the mean model $y = \mu(x) + \epsilon$ where

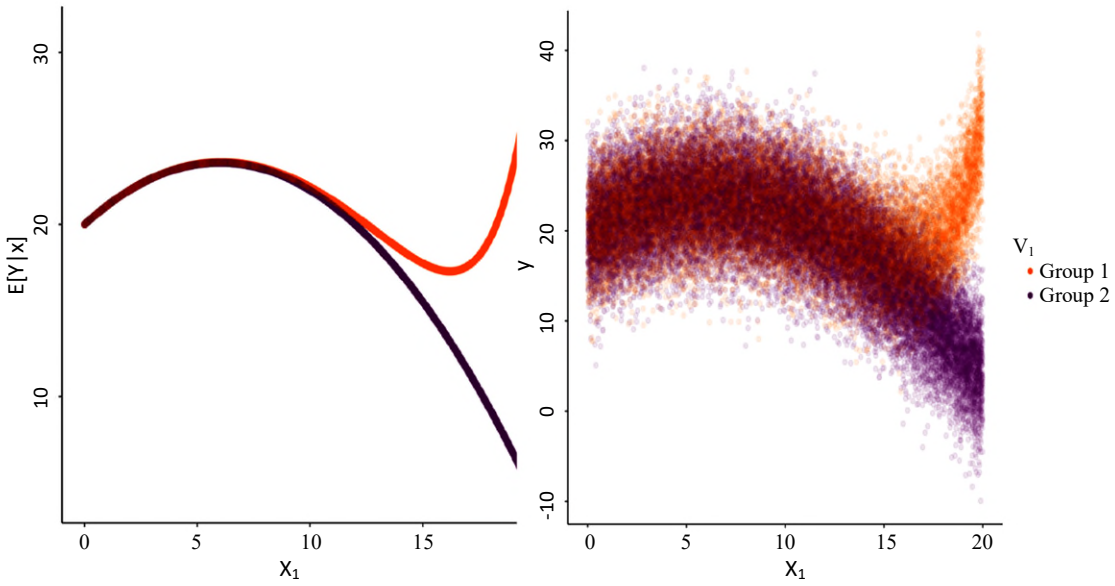
$$\mu(\mathbf{x}) = 0.2X_1(X_1 - 12) + 0.5\exp\{(X_1 - 15)\}\, \mathbb{I}_{\{V_1 \in \{A, B\}\}}.$$

The left side of Figure 4.2 shows the mean function, $\mu(\mathbf{x})$ and the right side shows a graph of the population values that were randomly generated from the model.

We also generate a variable Z, that we use for the size variable in order to test the methods under a PPS sample design. The values of the size variable Z are independently generated from the model $Z = \frac{1}{2}\mu(\mathbf{x}) + 5\eta$, where η has a chi-squared distribution with 5 degrees of freedom. The correlation between the size variable and Y is of 0.663, and so, in this example, the PPS design is informative.

For this simulation, we generate random finite populations with 1, 2, 4, 8, 16, and 32 thousand units. We then draw 500 repeated samples from each of these six finite populations. For each random sample we sample 5% of the units, which is 50, 100, 200, 400, 800, and 1,600 observations respectively. Again we use the sample data to estimate the forest model, use the forest model to predict the values of Y for the non-sampled units given the values of $\mathbf{X}$ in the population, and then use these values to estimate the population mean.

Figure 4.2 The values of $E[Y \mid \mathbf{x}]$ with respect to the variable X_1.



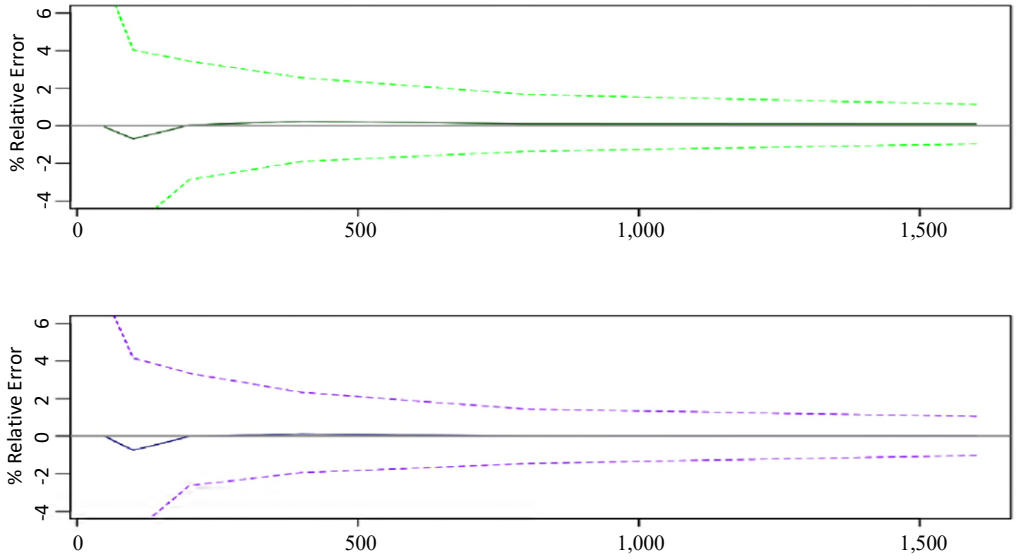
Note: The color denotes the values for the two groups of observations which is based on the value of the categorical variable V_1. Group 1 consists of observations where $V_1 \in \{A, B\}$ and Group 2 are all other observations.

By drawing random samples of increasing size and comparing our estimator to the true mean function, we evaluate the behavior of the accuracy and the variance of our estimator with respect to sample size. For instance, for a consistent estimator, one would expect the empirical confidence interval of average differences between the estimated values and values generated from the true mean function to contain zero. In addition, as the sample size increases, the variance of the average difference should decrease toward zero.

We can see that this is the case with the SRS design shown in Figure 4.3. shows the average over the 250 samples of the mean relative errors of the estimated population values as the sample size increases from 50 to 1,600. The distribution of the mean errors over the 500 repeated samples is centered right around 0 for every sample size for both forest modeling methods. In addition, the variance of the mean errors goes to zero as the sample size increases at about the same rate for both modeling methods.

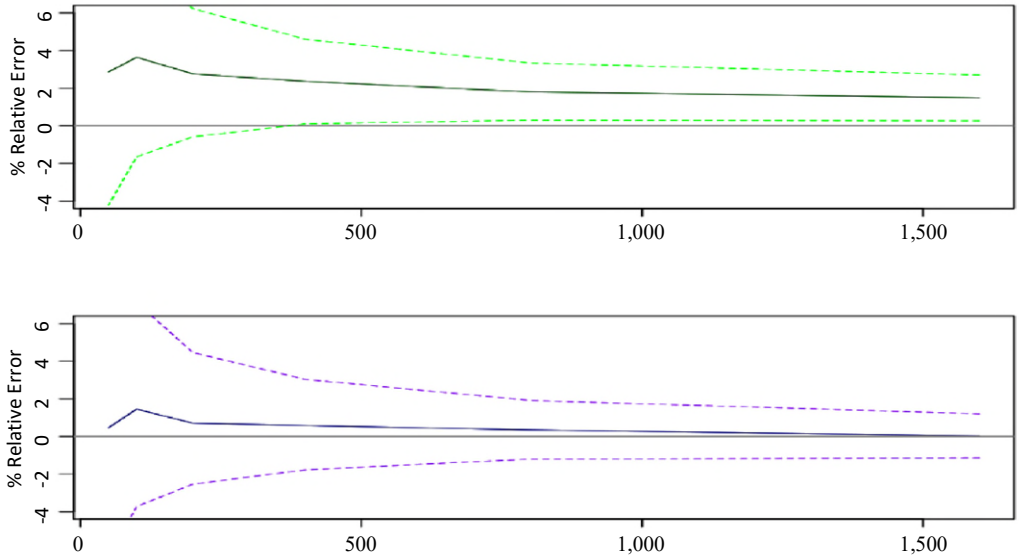
As we might expect, the story gets more interesting when the sample is drawn using a PPS design and the size variable is related to the variable of interest. In Figure 4.4, we see that the averages of the percent relative errors over the 250 repeated samples are longer as near zero for either method. Zero is contained within the middle 95% of values of percent relative errors when the sample size is below 800 for both methods because of the large variances in these values at small sample sizes. This interval no longer contains zero for the standard random forest algorithm as deviation of errors decrease with the increasing sample sizes. However, the proposed random forest method contains zero for every sample size for the proposed method.

Figure 4.3 Percent relative error by sample size for the regular (top) and the design consistent (bottom) random forest algorithms over repeated simple random samples.



Note: The solid line is the mean percent relative error over all the samples while the dashed lines give the 2.5 and the 97.5 percentile values.

Figure 4.4 Percent relative error by sample size for the regular (top) and the design consistent (bottom) random forest algorithms over repeated probability proportional to size samples.



Note: The solid line is the mean percent relative error over all the samples while the dashed lines give the 2.5 and the 97.5 percentile values.

These simulation results confirm the main result of the paper. That is, the proposed algorithm satisfying certain conditions that has been shown theoretically to be asymptotically design unbiased and consistent, has demonstrated over repeated samples to converge toward the true mean and be relatively unbiased.

5. Conclusions

Traditionally, complex survey data are collected to estimate finite population quantities. However, with the rise of machine learning methods, there is now greater interest in employing survey data in predictive problems and so the adaptation of machine learning methods to handle unequal probability data is now a vibrant area of research. In this paper, we present a new algorithm for estimating random forest models. This method which relies on independent random trees and a weighting procedure based on the weighted variability of the y-values is more appropriate for survey and other data collected from an informative design. We provide a set of conditions under which we show the method is design consistent for the conditional expectation of the variable of interest. The theoretical asymptotic unbiasedness and consistency of this algorithm is demonstrated through a simulation. The simulation studies are performed using real and generated data; we show that in practice the proposed method greatly reduces the bias of a random forest algorithm under an informative sample designs. In contrast, the estimates from the usual random forest method, which does not account for the sample design, are not unbiased under repeated informative samples. The estimates of both methods exhibited fairly similar mean squared errors. To ensure the individual trees are independent, our algorithm constructs truly random trees where a random variable and a random cut point are selected for each split.

Nalenz et al. (2024)'s approach of adjusting for an informative sample design is straight forward and interesting but it came out while this article was in review and so we have not compared it to our random trees method. Rather than avoid bootstrap sampling Nalenz et al. (2024) use a Hájek bootstrap. In their application this works very well as the outliers occur in the over sampled part of the population (units with low weights) and thus these units are down-weighted by the algorithm. However, in a general dataset, outlying values can also be associated with highly weighted survey units. Future work should compare the methods and, in the spirit of model aggregation, should consider blending the two approaches.

Acknowledgements

The authors would like to thank a lot of people later.

Appendix

Proofs and minor results

Proof of Corollary 2.1. Condition 1 requires the variable Y to have a finite fourth moment, then the random variable Y^2 has a finite second moment. Therefore, by taking Y^2 as the variable of interest in Proposition 2.1, we get a consistent estimator of $E_\xi[Y^2 \mid \mathbf{x}]$.

Lemma 6.1. *For a given tree j, if Conditions 1 through 7 are satisfied, then*

$$\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x}) < \infty, \text{ as } \nu \to \infty.$$

Proof of Lemma 6.1.

$$\begin{aligned}
\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} &= \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} \left(y_i - \tilde{\mu}_j(\mathbf{x})\right)^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \\
&= \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} \left(y_i^2 - 2y_i\tilde{\mu}_j(\mathbf{x}) + \tilde{\mu}_j^2(\mathbf{x})\right) \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \\
&= \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} y_i^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \\
&\quad - 2\tilde{\mu}_j(\mathbf{x}) \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} y_i \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} + \tilde{\mu}_j^2(\mathbf{x}) \\
&= \underbrace{\left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} y_i^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}}}_{I} \quad - \quad \underbrace{\tilde{\mu}_j^2(\mathbf{x})}_{II}
\end{aligned}$$

by equation (2.1).

By Proposition 2.1, $\tilde{\mu}_j(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}]$, so the term $II \to_p E_\xi^2[Y \mid \mathbf{x}]$ for each $j = 1 \ldots M$ and by Lemma 2.1,

$$I = \left(\tilde{\#}B_j^{n_\nu}(\mathbf{x})\right)^{-1} \sum_{i \in S} \pi_i^{-1} y_i^2 \mathbb{I}_{\{\mathbf{x}_i \in B_j^{n_\nu}(\mathbf{x})\}} \to_p E_\xi[Y^2 \mid \mathbf{x}],$$

for each $j = 1 \ldots M$. Therefore, $\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x})$, and by Condition 1, both the quantity I and II are finite with ξ probability 1.

Lemma 6.2. *If Conditions 1 through 7 are satisfied for each tree in a given forest with $M > 0$ trees, then*

$$\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$$

for all $\mathbf{x}$ *and all* $j = 1 \ldots M$, *as* $\nu \to \infty$.

Proof of Lemma 6.2. Because, for each $j = 1, \ldots, M$, the random variable $\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \geq 0$, by equation (3.3), the function

$$\lambda_j(\mathbf{x}) = \frac{(\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} + 1)^{-1}}{\sum_{j=1}^{M} (\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} + 1)^{-1}}$$

is continuous for each j. Also, by Lemma 6.1, $\tilde{V}_{B_j^{n_\nu}(\mathbf{x})} \to_p \mathrm{Var}(Y \mid \mathbf{x})$ for each j. Therefore, by the Continuous Mapping Theorem,

$$\lambda_j(\mathbf{x}) \to_p \frac{(\mathrm{Var}(Y \mid \mathbf{x})) + 1)^{-1}}{\sum_{j=1}^{M} (\mathrm{Var}(Y \mid \mathbf{x}) + 1)^{-1}} = \frac{1}{M}.$$

This lemma states that because the variance converge to $V[Y \mid \mathbf{x}]$ and so all end-node weights converge to $1/M$ asymptotically. However, note that for finite n, the weights will differ substantially depending on the efficiency gain from the random splits resulting in the end-node. This is the adaptiveness that has been added to the procedure and where all the real work is being done.

Lemma 6.3. *For a fixed $M > 0$, if Conditions 1 through 7 are satisfied for each tree in the forest of M trees, then*

$$M\widehat{\text{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p 0,$$

for all $\mathbf{x}$ *as* $\nu \to \infty$.

Proof of Lemma 6.3. Since $\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ by Lemma 6.2, $\bar{\lambda} = M^{-1}\sum_{j=1}^{M}\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ by the Continuous Mapping Theorem. Likewise, $\bar{\mu} = M^{-1}\sum_{j=1}^{M}\tilde{\mu}_j(\mathbf{x}) \to_p E[Y \mid \mathbf{x}]$ since each $\tilde{\mu}_j(\mathbf{x}) \to_p E[Y \mid \mathbf{x}]$ by Proposition 2.1. Therefore, each of the terms of equation (3.7),

$$(\lambda_j(\mathbf{x}) - \bar{\lambda})(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}) \to_p 0.$$

Once again applying the Continuous Mapping Theorem to

$$M\widehat{\text{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) = \frac{1}{(1 - M^{-1})}\sum_{j=1}^{M}(\lambda_j(\mathbf{x}) - \bar{\lambda})(\tilde{\mu}_j(\mathbf{x}) - \bar{\mu}),$$

gives the result.

Proof of Proposition 3.1. Since each $\lambda_j(\mathbf{x}) \to_p \frac{1}{M}$ by Lemma 6.2, $\tilde{\mu}_j(\mathbf{x}) \to_p E_\xi[Y \mid \mathbf{x}]$ by Proposition 2.1 and $M\widehat{\text{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p 0$, by Lemma 6.3, we can apply the Continuous Mapping Theorem to get that

$$\mathcal{F}_{n_v}(\mathbf{x}) = \sum_{j=1}^{M}\lambda_j(\mathbf{x})\,\tilde{\mu}_j(\mathbf{x}) - M\widehat{\text{cov}}(\lambda(\mathbf{x}), \tilde{\mu}(\mathbf{x})) \to_p \sum_{j=1}^{M}\frac{1}{M}E_\xi[Y \mid \mathbf{x}] + 0 = E_\xi[Y \mid \mathbf{x}].$$

References

Arlot, S., and Genuer, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.

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

Bilton, P., Jones, G., Ganesh, S. and Haslett, S. (2017). Classification trees for poverty mapping. *Computational Statistics & Data Analysis*, 115, 53-66.

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Buskirk, T.D. (2018). Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Surv Pract*, 11, 2709.

Dagdoug, M., Goga, C. and Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 1-18.

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

Gajowniczek, K., Grzegorczyk, I., Ząbkowski, T. and Bajaj, C. (2020). Weighted random forests to improve arrhythmia classification. *Electronics*, 9(1), 99.

Gelman, A., King, G. and Liu, C. (1998). Not asked and not answered: Multiple imputation for multiple surveys. *Journal of the American Statistical Association*, 93(443), 846-857.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5, 361-74.

Hong, H.G., and He, X. (2010). Prediction of functional status for the elderly based on a new ordinal regression model. *Journal of the American Statistical Association*, 105(491), 930-941.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Krebs, M.A., Reeves, M.C. and Baggett, L.S. (2019). Predicting understory vegetation structure in selected western forests of the United States using fia inventory data. *Forest Ecology and Management*, 448, 509-527.

Kshirsagar, V., Wieczorek, J., Ramanathan, S. and Wells, R. (2017). Household poverty classification in data-scarce environments: A machine learning approach. *arXiv preprint arXiv:1711.06813*, 2017.

Lavallée, P., and Beaumont, J.-F. (2015). Why we should put some weight on weights. *Survey Methods: Insights from the Field (SMIF)*.

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3), 18-22. URL: https//CRAN.R-project.org/doc/Rnews/.

Little, R.J. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466), 546-556.

Loh, W.-Y. (2008). Classification and regression tree methods. *Encyclopedia of Statistics in Quality and Reliability*, 1, 315-323.

Lumley, T. (2020). survey: analysis of complex survey samples, 2020. R package version 4.0.

Mashreghi, Z., Haziza, D. and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.

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

Morgan, J.N., and Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302), 415-434.

Nalenz, M., Rodemann, J. and Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine Learning*, 113(6), 3379-3398.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, 317-337.

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

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

Shahhosseini, M., and Hu, G. (2020). Improved weighted random forest for classification problems. *International Online Conference on Intelligent Decision Science*, 42-56. Springer.

Toth, D. (2024). *rpms: Recursive Partitioning for Modeling Survey Data*. R package version 1.0.0.

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

Wieczorek, J. (2023). Design-based conformal prediction. *Survey Methodology*, 49, 2, 443-473. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2023002/article/00007-eng.pdf.

Williams, J.K., Neilley, P.P., Koval, J.P. and McDonald, J. (2016). Adaptable regression method for ensemble consensus forecasting. *Thirtieth AAAI Conference on Artificial Intelligence*.

Winham, S.J., Freimuth, R.R. and Biernacka, J.M. (2013). A weighted random forests approach to improve predictive performance. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6), 496-505.

Yang, D.K., and Toth, D.S. (2022). Analyzing the association of objective burden measures to perceived burden with regression trees. *Journal of Official Statistics*, 38(4), 1125-1144.