Erasmus School of Economics

Bachelor Thesis (Econometrie en Operationale Research, Quantitative Finance)

Treatment Effect Estimation:

A performance comparison between Random Forests and Bayesian Additive Regression Trees

Name student:
Catelijn, M.B.
Student number:
494933mc

Supervisor:
Klooster, J.
Second Assessor:
O'Neill, E.P.

Date:
July 4, 2021

Abstract

Wager and Athey (2018) have derived statistical inference for random forests and shown that these forests outperform the k-nearest neighbours algorithm. In this paper we will verify these results and extend their research by comparing these forests to a Bayesian Additive Regression Trees (BART) model, in a simulation and observational setting. In the simulation setting, we find that in general BART outperforms the random forests. In the observational setting, BART and the random forests provide similar estimations. In practice, we advice to use BART over the random forests.

Keywords: Treatment Effect - Random Forest - BART

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

In practice, research in various fields, such as health, politics and economics, is causal of nature (Pearl, 2010). This research often boils down to the same idea: 'How does a certain treatment affect a certain individual?'. This can then be extended to addressing the heterogeneity of these effects. Namely, how does the treatment affect different groups of people? For example, what characteristics are important for a new drug to perform best? What is the best time to broadcast an ad on which platform? These are naturally impactful questions to answer in practice.

Let us look at a recent example, during the current COVID-19 crisis there was a race to develop a vaccine. When treated with this vaccine the patient should gain some sort of protection. Either in the form of immunity or reduction of symptoms. However, it might be possible that the vaccine works better for people with different characteristics, for example, it is best at protecting elderly people, and worse for others. Since the race to the vaccine was a time-sensitive production process, the first producer will receive a lot of orders, producers might have an incentive to report only the most favourable statistics. Assmann et al. (2000)) show that this is not uncommon in clinical trials, people might look for subgroups in the data with very high treatment effects, to then only report the results of that group. Therefore, researchers have to predefine the statistical analysis to prevent illegitimate trial results.

In this paper, we will replicate and extend the research performed in the influential paper of Wager and Athey (2018). Various advanced non-parametric machine learning and deep learning methods, such as neural networks and random forests, are often referred to as black-box models. Meaning that the interactions and the interpretations of the model are not directly observed. Therefore classical statistical inference is not possible. Recently, Wager and Athey (2018) derived asymptotics of the estimate this inference is now possible. Allowing to perform model diagnostics and to create confidence intervals around estimates. These machine learning methods generally tend to outperform traditional econometric models in terms of point estimates. Since interval estimates obtain more information than point estimates this derivation is of great value.

Wager and Athey (2018) provide extensive statistical background leading up to their main finding of growing a random causal forest that is capable of providing consistent and asymptotically normal estimates for heterogeneous treatment effects. Furthermore, they perform a simulation experiment which shows that their causal random forests have better performance compared to a baseline method of k-nearest neighbour estimation. These promising results sparked the idea of extending the current literature by applying the methodology presented by Wager and Athey (2018) on preceding research in the estimation of heterogeneous treatment effects based on observational data and comparing it to a competitive model. More concisely:

RQ: *How do causal random forests perform compared to Bayesian Additive Regression Trees (BART) when exposed to observational data?*

To answer this research question, we will first describe the methodology presented by Wager and Athey (2018) and define the Bayesian Additive Regression Trees (BART) model. After which we will replicate their simulation experiment to verify the validity of the performance of the causal random forest and compare this to the performance of BART. We find that in general the BART outperforms the random forest. Finally, we will apply both the causal random forest and BART to observational data of Green and Kern (2012) and compare their performance. The dataset contains data on respondents of The General Social Survey (GSS), they were asked the following question: "... are we spending too little, too much or about the right amount on ...?" where for some respondents the final word was **welfare** and for the others, the final words were **assistance to the poor**. We find that the random forest and BART provide similar estimations, however, BART seems to be more sensitive to changes in the dependant variables. After this comparison we will provide some concluding remarks as well as suggest implications for both the academic world and in practice.

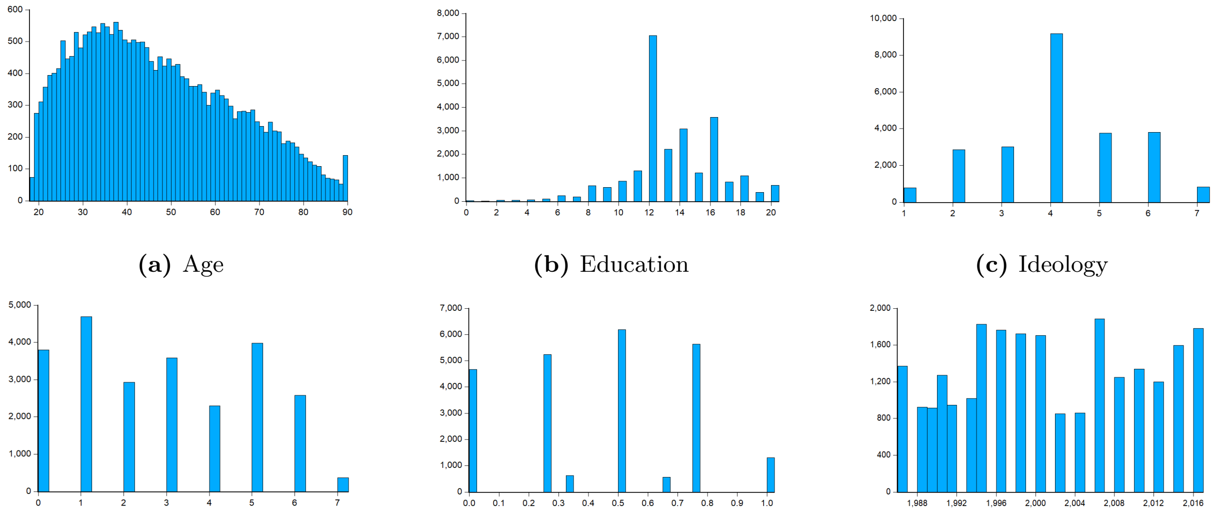
In Section 2 we will describe the dataset for the observational study. Then, in Section 3 we will define all the models and describe the simulation set-up. After which, we describe the results of the simulation and the observational study in Section 4. Finally, in Section 5 we provide some concluding remarks.

2 Data

In this paper, we analyse the performance of a random forest model in a simulation setting and in an observational setting. Due to the mathematical specificity of the simulation setting, this will be discussed at the end of the methodology section. In this section, we will describe our observational setting.

The observational setting is inspired by Green and Kern (2012). In their experimental setup, they estimate the conditional average treatment effects for the change of wording of a survey question regarding welfare. Respondents of The General Social Survey (GSS) were asked the following question: "... are we spending too little, too much or about the right amount on ...?" where for some respondants the final word was **welfare** and for the others the final words were **assistance to the poor**. Where welfare is considered the treated group and assistance to the poor is considered the control group.

The GSS is conducted in the United States of America. It is a substantial survey about various social statements. We will analyse datasets over the period of 1986 until 2016, containing a total of 24209 observations. The dataset contains variables on demographic information, political ideology and racist believes. Figure 1 displays density histograms for all the variables. *Age* is the age in years of the respondent with mean 46.29. Note that there is an irregular spike for Age at a value of 89, this is due to the fact that respondents could only answer ‘89 or older’, therefore, all the older ages are truncated to 89. Year is the year in which the GSS survey was conducted. *Education* is the education in years of the respondent, with an average of 13.36. The political party of the respondent is captured on a 7 point scale in the variable *Party*, where 1 is ‘Strong democrat’, 4 is ‘Independent’ and 7 is ‘Strong republican’. If the respondent answered with ‘Other party’ this is given the value 0. *Ideology* is a variable on a 7 point scale which contains the political ideology, ranging from ‘Extremely liberal’ as 1 to ‘Extremely conservative’ as 7, with mean 4.11. *Racist* is a variable ranging from 0 to 1 where non racist is classified by 0 and racist by 1, with mean 0.38. This variable was computed by averaging answers, where the most racist answer was coded as 1, of 4 yes-or-no questions about inequality due to race. When one or more answers are missing the average is computed over the remaining answers. An overview of the precise wording of the questions can be found in Appendix A.



(a) Age

(b) Education

(c) Ideology

(d) Party

(e) Racist

(f) Year

Figure 1: Distribution of variables

3 Methodology

In this section we will first describe the setting of treatment effect estimation along with the specification of the random forest model. Second, we show that these forests produce asymptotically normal estimates. Thirdly, we introduce the BART model. After which we compare the random forest and the BART. Finally, the experimental setup is defined.

We define the treatment effect as

$$\tau(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \mid X_i = x\right], \tag{1}$$

where $Y_i^{(1)}$ and $Y_i^{(0)}$ are possible outcomes for individual i when experienced treatment or not experienced treatment respectively[1], and X_i denotes the feature space for individual i. However, as noted in the introduction, we can only observe one of the possible outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$. Therefore it is not possible to compute $Y_i^{(1)} - Y_i^{(0)}$ to directly estimate $\tau(x)$. If we define W_i as a binary treatment variable, taking a value of 1 if individual i has experienced treatment and 0 otherwise. We can then define the treatment outcome Y_i as

$$Y_i = \begin{cases} Y_i^{(1)}, & \text{if } W_i = 1; \\ Y_i^{(0)}, & \text{if } W_i = 0. \end{cases} \tag{2}$$

Then the assumption of unconfoundness, the potential outcomes for Y_i are independent of the treatment assignment conditional on the features X_i, denoted by:

$$\left\{Y_i^{(1)}, Y_i^{(0)}\right\} \perp\!\!\!\perp W_i \mid X_i. \tag{3}$$

Rosenbaum and Rubin (1983) found that when this assumption holds, equation 1 can be rewritten as:

$$\tau(x) = \mathbb{E}\left[Y_i\left(\frac{W_i}{e(x)} - \frac{1 - W_i}{1 - e(x)}\right) \mid X_i = x\right], \tag{4}$$

where

$$e(x) = \mathbb{E}\left[W_i \mid X_i = x\right], \tag{5}$$

is the propensity score. The expectation over a binary variable can be interpreted as probability. Therefore, the propensity score $e(x)$ can be interpreted as the probability that an individual experiences treatment given its feature set X_i. Consequently, given equation 4, estimating $\tau(x)$ reduces to estimating $e(x)$. This is the root of various methods for causal inference. However, Wager and Athey (2018) uses an approach to consistently estimate the treatment effect $\tau(x)$, using propensity trees and the assumption of unconfoundness, without directly estimating $e(x)$. This approach will now be described.

Wager and Athey (2018) show that they have build trees that are capable of achieving consistency and asymptotic normality. These trees are based of the Classification and Regression Trees (CART) (Breiman et al., 1984). This is a machine learning method that recursively partitions observations based on a "splitting criterion". The algorithm stops once the entire set of observations,

[1] In the literature, this is often referred to as 'control'. We will therefore refer to $Y_i^{(0)}$ as the control outcome.

containing (Y_i, X_i) pairs, is divided in "leaves" L, containing a small enough amount of observations. By construction these observations are very similar. Therefore, for a new observation x we evaluate the prediction $\bar{Y}(x)$ by finding the leaf $L(x)$ and analysing

$$\bar{Y}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{\{i:X_i \in L(x)\}} Y_i. \tag{6}$$

This method is believed to perform well, if the assumption holds that within each leaf, observations are approximately identically distributed. Such that they seem to all be outcomes of the same random experiment.

This idea of the CART regression tree can be converted to a causal tree. We want to quantify the effect of a treatment. Therefore, instead of looking at pairs of (Y_i, X_i) observations we will now consider (Y_i, W_i). Following the same procedure as described before we can now estimate the treatment effect $\tau(x)$ for any new observation x in leaf $L(x)$ as

$$\begin{aligned}\widehat{\tau}(x) = \frac{1}{|\{i : W_i = 1, X_i \in L(x)\}|} \sum_{\{i:W_i=1, X_i \in L(x)\}} Y_i - \\ \frac{1}{|\{i : W_i = 0, X_i \in L(x)\}|} \sum_{\{i:W_i=0, X_i \in L(x)\}} Y_i \\ = \bar{Y}^{(1)}(x) - \bar{Y}^{(0)}(x).\end{aligned} \tag{7}$$

Equation 7 shows that the treatment effect $\widehat{\tau}(x)$ can be estimated by first determining which leaf x belongs to, then computing the mean treated outcome minus the mean control outcome within its leaf. Combining multiple of these causal trees creates a causal forest that is consistent for $\tau(x)$. This forest can be grown by taking an ensemble of T causal trees, that all provide individual estimates $\widehat{\tau}_t(x)$, as: $\widehat{\tau}(x) = T^{-1} \sum_{t=1}^{T} \hat{\tau}_t(x)$.

These forests are, as shown by Wager and Athey (2018), consistent and asymptotically normal once the assumption of honesty holds. This assumption is defined as follows (Wager and Athey, 2018): "a tree is honest if, for each training example i, it only uses the response Y_i to estimate the within-leaf treatment effect τ using 7 or to decide where to place the splits, but not both."

So in order to be able to apply the theoretical findings of Wager and Athey (2018), that guarantee consistency and asymptotic normality, we need a procedure to create these honest trees that combined create a honest forest. Two procedures are suggested by Wager and Athey (2018) which we will directly quote for reasons of accuracy.

Procedure 1. Double-Sample Trees Double-sample trees split the available training data into two parts: one half for estimating the desired response inside each leaf, and another half for placing splits. Input: n training examples of the form (X_i, Y_i) for regression trees or (X_i, Y_i,W_i) for causal trees, where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random subsample of size s from $\{1, ..., n\}$ without replacement, and then divide it into two disjoint sets of size $|\mathcal{I}| = s/2$ and $|\mathcal{J}| = s/2$
2. Grow a tree via recursive partitioning. The splits are chosen using any data from the $\mathcal{J}$ sample and X- or W-observations from the $\mathcal{I}$- sample, but without using Y-observations from the $\mathcal{I}$-sample.
3. Estimate leafwise responses using only the $\mathcal{I}$-sample observations

The assumption of honesty requires that for every observation i the outcome Y_i is only used in either the placement of the splits within the tree or for the estimation of the treatment effect within the leaf. Procedure 1 splits a subsample into two disjoint sets $\mathcal{I}$ and $\mathcal{J}$. The latter is solely used for the determination of the splits. Furthermore, all the information in $\mathcal{I}$ *except* the outcome Y_i is also used for placing splits. Finally only the outcome variables Y from $\mathcal{I}$ are used for the estimation of the treatment effect. Therefore, it is clear that the honesty assumption holds for double-sample trees.

Procedure 2. Propensity Trees Propensity trees use only the treatment assignment indicator W_i to place splits, and save the responses Y_i for estimating τ. Input: n training examples (X_i, Y_i, W_i), where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random subsample $\mathcal{I} \in \{1, ..., n\}$ of size $|\mathcal{I}| = s$ (no replacement).
2. Train a classification tree using sample $\mathcal{I}$ where the outcome is the treatment assignment, that is, on the (X_i, W_i) pairs with $i \in \mathcal{I}$. Each leaf of the tree must have k or more observations of each treatment class.
3. Estimate $\tau(x)$ using 7 on the leaf containing x.

In step 2, the splits are chosen by optimizing, for example, the Gini criterion used by CART for classification.(Breiman et al., 1984)

Since procedure 2 does not use any Y_i observations for placing the splits and uses these solely for estimation of the treatment effect it immediately follows that the honesty assumption holds.

3.1 Consistency

The first main theoretical finding is that Wager and Athey (2018) found their random forests to be consistent as long as two more assumptions, besides honesty, hold. The exact mathematical argument as to why these assumptions have to hold in order to attain consistency is beyond the scope of this paper, but can be found in Theorem 3.1 of Wager and Athey (2018). However, it is of importance to know what assumptions lay the foundations for the model. Therefore, we will intuitively explain the two assumptions needed for consistency. First, we need to assume that both $\mathbb{E}\left[Y_i^{(1)} \mid X_i = x\right]$ and $\mathbb{E}\left[Y_i^{(0)} \mid X_i = x\right]$, are Lipschitz continuous. Lipschitz continuity limits the absolute value of the slope of the functions by a certain Lipschitz constant. This implies that all the derivatives of our outcome variable Y conditioned on X will not surpass a certain magnitude. Second, we need an overlap assumption. This assumption implies that there are sufficient treated and non-treated observations close to each observation x. Once these assumptions hold the causal random forest attains consistency. For large enough datasets we expect these assumptions to both hold.

3.2 Asymptotic normality

The second main theoretical finding of Wager and Athey (2018) is that the estimated treatment effect $\widehat{\tau}(x)$ is assymptotically normally distributed. Namely, they found that

$$\frac{\widehat{\tau}(x) - \tau(x)}{\sigma_n(x)} \to \mathcal{N}(0,1), \qquad \text{for a sequence } \sigma_n(x) \to 0. \tag{8}$$

This means that, as long as the sampling standard deviation approaches zero as the sample size n increases, the estimated treatment effect follows a normal distribution. Furthermore, Equation 8 only holds when the assumptions of consistency hold, as well that the sub-sample size s is in the same order of magnitude as $s \asymp n^\beta$ for some $\beta_{\min} < \beta < 1$. Here $\beta_{\min}$[2] is a computable quantity which increases for higher dimensional problems. This implies that β has a higher upperbound for higher dimensional problems, which would mean that the sub-sample size s is allowed to approach the sample size n.

In addition, Wager and Athey (2018) find a method to consistently estimate the variance $\sigma_n^2(x)$. This estimator, called the inifinitesimal jackknife variance estimator, is similar to a standard variance estimator, however it uses a first order Taylor expansion around the true loss function in order to be computationally more efficient. It is based on the original work of Jaeckel (1972). The variance estimator is defined as

[2] Equation 14 (Wager and Athey, 2018)

$$\widehat{V}_{\mathrm{IJ}}(x) = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}_*[\widehat{\tau}_t^*(x), N_{it}^*]^2. \tag{9}$$

Here, $\widehat{\tau}_t^*(x)$ is the treatment effect estimate provided by tree t and $N_{it}^* \in \{0,1\}$ indicates whether the i'th observation was used in the estimation of the t'th tree. The stars indicate that the covariance is taken over the entire set of trees $t = 1, \ldots, T$ in the forest. This estimate is consistent such that the estimate converges in probability to the sampling variance as

$$\frac{\widehat{V}_{\mathrm{IJ}}(x)}{\sigma_n^2(x)} \to_p 1. \tag{10}$$

Since we are now able to estimate the treatment effect as $\widehat{\tau}_t(x)$ and the sample standard deviation $\sigma_n(x)$ via the infinitesimal jackknife estimate as $\sqrt{\widehat{V}_{\mathrm{IJ}}(x)}$, we can now compute confidence intervals for the estimated treatment effect $\widehat{\tau}_t(x)$. These characteristics allow us to use classical statistical inference on these non-parametric honest random forests.

3.3 Bayesian Additive Regression Trees (BART)

As an extension on the research of Wager and Athey (2018) we will apply the random forest algorithm on observational data and we will compare the random forest to a more competitive model. We will apply the forest to the observational data described in the data. The random forest model will be compared to a Bayesian Additive Regression Trees model (BART). BART was used by Green and Kern (2012) to estimate conditional average treatment effects (CATEs) in the dataset we will also analyse. In order to compare performance of the random forest to BART on observational data it is reasonable to first compare performance in a controlled environment. Therefore, we will apply BART to the simulation experiments. In this section we will define BART.

BART is a "sum-of-trees model", first developed by Chipman et al. (2010). It works similar to the random forests as it combines estimates of various trees into one final estimate. Let us first describe some data generating process as

$$Y = f(x) + \varepsilon, \qquad\qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{11}$$

where $f(\cdot)$ is some unknown function. We will approximate this function $f(\cdot)$ by a sum-of-trees model $h(\cdot)$ which is a sum of 'smaller' functions $g(\cdot)$ as

$$Y = f(x) + \varepsilon \approx h(x) + \varepsilon = \sum_{j=1}^{m} g(x; T_j, M_j) + \varepsilon, \qquad\qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{12}$$

where T_j denotes a binary decision tree containing a set of decision rules and a set of terminal nodes with cardinality b. The set of values of each of the terminal nodes is denoted as M_j, with

$M = \{\mu_1, \ldots, \mu_b\}$. The function $g(\cdot)$ appoints a value $\mu_i \in M$ based on characteristics x. The conditional mean of Y given the characteristics x for the sum-of-trees model $\mathbb{E}[Y|x]$ is the sum of the terminal node values of m trees, $\sum_{j=1}^{m} \mu_{ij}$, where μ_{ij} is the terminal node value of the j'th tree, such that each individual tree contributes partly to the total conditional mean $\mathbb{E}[Y|x]$. In order to complete the initialization of the BART model we need to implement the Bayesian aspect to the sum-of-trees model. In order to achieve this, we will introduce the basics of Bayesian statistics, such that even without extensive knowledge on the topic the BART model can be understood.

Bayesian Statistics

Bayesian statistics is a view on statistics based on Bayes' theorem. In traditional statistics parameters are viewed as unknown to-be-estimated quantities. However, Bayesian statistics views these parameters not as fixed quantities but as unknown random variables. This different perspective requires a different estimation approach. In order to explain this estimation approach we will first introduce some required terminology. The **prior** distribution is a probability distribution appointed to the unknown parameters θ as $p(\theta)$. The distribution of conditioning the parameters θ on observational data X_n is called the **sample** distribution, denoted as $p(X_n|\theta)$, note that X_n is a sample containing n observations. The **posterior** distribution is a probability distribution, denoted as $p(\theta|X_n)$. The posterior is obtained when updating the prior with the sample using Bayes' theorem. Bayes' theorem is defined as

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}. \tag{13}$$

In the framework of Bayesian inference this translates to

$$p(\theta|X_n) = \frac{p(X_n|\theta)p(\theta)}{p(X_n)} \propto p(X_n|\theta)p(\theta), \tag{14}$$

where $p(X_n)$ can be considered a normalizing constant which makes sure that the posterior is scaled to be probability density function[3].

We can now translate this theory to the BART model. Note that BART is a non-parametric approach, therefore our choice of θ is not a set of parameters but it is the structure of the sum-of-trees model. Therefore, we need to determine on what aspect we want to induce a prior. The entire structure of the sum-of-trees model can be determined by (i) the set of decision rules and terminal values encapsuled in T_j, (ii) the values of the terminal nodes captured in M_j, (iii) and the standard deviation of the error term denotes as σ. (i) and (ii) can both be captured for the j'th tree as one combination of (T_j, M_j), with $j = 1, ..., m$. Accordingly, we need to compute a prior containing these three aspects. As Chipman et al. (2010) suggests, we will look for priors that

[3] Guaranteeing that the integral over the posterior is 1.

assume independence, such that

$$p((T_1, M_1), ..., (T_m, M_m), \sigma) = \left[\prod_j p(T_j, M_j)\right] p(\sigma) = \left[\prod_j p(M_j|T_j)p(T_j)\right] p(\sigma), \tag{15}$$

where,

$$p(M_j|T_j) = \prod_i p(\mu_{ij}|T_j). \tag{16}$$

Therefore, we need to define a tree prior (a), a terminal node prior (b) and a σ prior (c), which we denote as $p(T_j)$, $p(\mu_{ij}|T_j)$ and $p(\sigma)$ respectively.

(a) Following Chipman et al. (2010) the tree prior, $p(T_j)$, is specified by three elements. First, the prior for the probability that at depth $d = 0, 1, 2, ...$ a node is nonterminal is given by

$$\alpha(1+d)^{-\beta}, \qquad \alpha \in (0,1), \beta \in [0, \infty), \tag{17}$$

α and β are hyperparameters which can be chosen to determine the shape of the above described function. Chipman et al. (2010) suggests setting $\alpha = 0.95$ and $\beta = 2$. This choice of hyperparameters makes it such that the probability of having a 'deep' tree is very unlikely. Consequently, individual trees in the sum-of-trees model are usually 'shallow' trees, containing only a few node layers. As illustration, This construction implicitly prevents overfitting[4] to a certain extent.

Second, at each interior node the splitting variable assignments are determined by a uniform prior distribution. Third, the splitting rules at each interior node are given an uniform prior distribution on the discrete set of available splitting values.

(b) The prior for μ_{ij} condtioning on T_j, $p(\mu_{ij}|T_j)$, is a conjugate prior given by

$$\mu_{ij} \sim \mathcal{N}(0, \sigma_\mu) \qquad \text{where } \sigma_\mu = \frac{0.5}{k\sqrt{m}} \tag{18}$$

where m is the number of trees and k is a parameter to be chosen, Chipman et al. (2010) found that values between 1 and 3 yield good results. A conjugate prior is a prior distribution that causes the posterior function to be off the same 'family'. In our case, since we chose a normal distribution as our conjugate distribution this implies that our posterior distribution is a Gaussian distribution. This prior guarantees that, for a BART model containing a large amount of trees m, the σ_μ of each individual tree is small effectively diminishing the impact of a single tree in the model.

(c) The prior for the standard deviation of the error term σ, $p(\sigma)$ is set to be an inverse chi-squared distribution with two hyperparameters ν and λ. Such that, $\sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}$, where ν is the amount of degrees of freedom and λ is the scale. Chipman et al. (2010) suggest a default setting of $\nu = 3$ and λ is determined based on a rough estimate of the standard deviation $\widehat{\sigma}$, for example

[4] The process that non-parametric methods capture patterns of a certain dataset, instead of the patterns of the underlying data generating process.

by computing the sample standard deviation. Then select a λ such that the q'th quantile of the quantile is located at this rough estimate, $\mathbb{P}[\sigma < \widehat{\sigma}] = q$, with a recommended value of q as 0.90. These settings makes it so that there is not an extreme amount of probability mass around small values of σ. Since small values of σ enables the model to overfit, this setting counters overfitting. The posterior distribution of BART given observed data x is defined as

$$p((T_1, M_1), ..., (T_m, M_m), \sigma | x). \tag{19}$$

In order to compute this posterior Chipman et al. (2010) propose a MCMC backfitting algorithm, which is in general terms a Gibbs sampler.[5] In order to compute estimates of the treatment effect $\widehat{\tau}(x)$, the backfitting algorithm will first construct the posterior distribution after which we take S samples $\tau_s(x)$ from this distribution. Then our estimate will be the mean of these samples, that is $\widehat{\tau}(x) = S^{-1} \sum_{s=1}^{S} \hat{\tau}_s(x)$. Since the mathematical specification of the posterior distribution cannot be directly derived we cannot suggest an estimator of the variance of the estimator. However, we can still compute confidence intervals around $\widehat{\tau}(x)$ by calculating the quantiles over the sample for any level of significance.

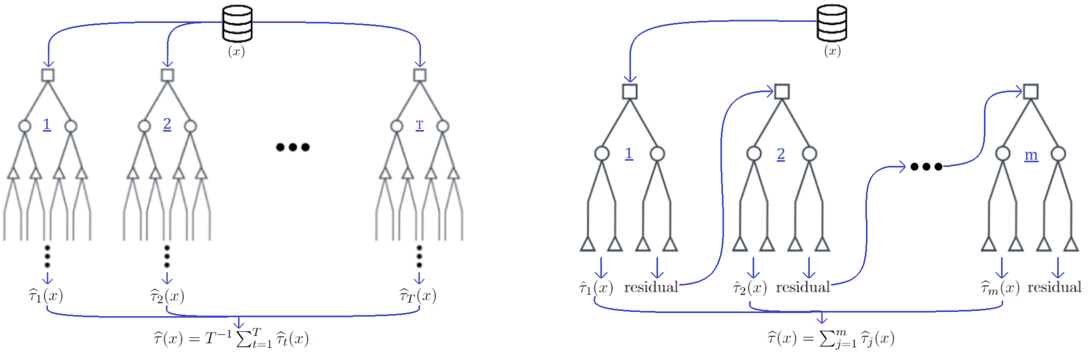
3.4 Comparison of random forest and BART

Since the non-parametric approaches used in this paper both look and sound very similar, for example both using trees for estimation, we want dedicate some attention to a methodological comparison of the random forest and BART. This might allow for better understanding of the difference in performance of the models. This comparison is displayed graphically in Figure 2.

First of all, both models use decision trees as tool to partition the feature space. However, BART restricts the size of these trees by setting a restrictive prior distribution on the depth of the tree, resulting in shallow trees. Whereas the random forest does not have such a constraint, allowing for deeper trees.

Secondly, both models use multiple of these trees in their final estimation. Nonetheless, the models differ in the way they combine the individual trees. The random forest contains T trees, where each tree t, partitions the data x, to eventually compute the estimated treatment effect $\widehat{\tau}_t(x)$. The final estimate for the random forest can be calculated by taking the mean of all the individual trees. In contrast, trees in the BART model ”pass on” their contribution. The first tree takes as input the data x and all the priors defined in the previous subsection. Since the trees are shallow, the first tree will slightly update its priors based on the believes, provided in the data. This makes it so the first tree captures a part of $\tau(x)$. Then it passes on the data that it has not yet captured to the next tree in the form of a residual. This process continues until the last tree. The estimated treatment effect can the be computed by summing over all the fits of the individual trees.

[5] The technical description of this process is beyond the scope of this paper. It can be found in Section 3 of Chipman et al. (2010).


(x)
1
2
T
$\widehat{\tau}_1(x)$
$\widehat{\tau}_2(x)$
$\widehat{\tau}_T(x)$
$\widehat{\tau}(x) = T^{-1}\sum_{t=1}^{T}\widehat{\tau}_t(x)$
(x)
1
2
m
$\widehat{\tau}_1(x)$ residual
$\widehat{\tau}_2(x)$ residual
$\widehat{\tau}_m(x)$ residual
$\widehat{\tau}(x) = \sum_{j=1}^{m}\widehat{\tau}_j(x)$

(a) Random forest

(b) BART

Figure 2: Graphical overview of the models

The left and right panels graphically display how respectively the random forest and BART produce their estimates. The random forest creates deep, uncorrelated, trees to then ensemble into an estimate. In contrast, each tree in the BART model explains a fraction of the real effect. This is then summed up and captured as the final estimate.

3.5 Simulation experiments

The simulation study of Wager and Athey (2018) aims to address two types of bias often present in practice. First, areas of relative stable $\tau(x)$ need to be identified. Second, no bias should be created by varying the propensity scores $e(x)$. To address this they perform to simulation experiments using a k-nearest neighbors (k-NN) method as baseline

$$\widehat{\tau}_{\text{KNN}}(x) = \frac{1}{l}\sum_{i \in S_1(x)} Y_i - \frac{1}{k}\sum_{i \in S_0(x)} Y_i, \tag{20}$$

where $S_1(x)$ and $S_0(x)$ are sets containing the k nearest neighbors for x in treatment group and control group respectively. The k-NN estimates $\widehat{\tau}_{KNN}(x)$ are assumed to be normally distributed

$$\frac{\widehat{\tau}_{\text{KNN}}(x) - \tau(x)}{\widehat{V}\left(\widehat{\tau}_{\text{KNN}}(x)\right)} \to \mathcal{N}(0, 1), \tag{21}$$

where

$$\widehat{V}\left(\widehat{\tau}_{\text{KNN}}(x)\right) = \frac{\widehat{V}(S_1) + \widehat{V}(S_0)}{k(k-1)}, \tag{22}$$

with $\widehat{V}(S_i)$ the sample variance of S_i with $i = 0, 1$.

The overall setup is as follows, let n and d denote the sample size and dimension of the feature space respectively. The following functions are defined:

$$\text{main effect: } m(x) = 0.5 \times \mathbb{E}\left[Y^{(0)} + Y^{(1)} \mid X = x\right], \tag{23}$$

$$\text{treatment effect: } \tau(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \mid X = x\right], \tag{24}$$

$$\text{treatment propensity: } e(x) = \mathbb{P}\left[W = 1 \mid X = x\right]. \tag{25}$$

In all experiments unconfoundness is assumed, X is uniformally distributed $X \sim U([0,1]^d)$ and homoskedastic white noise is present $Y^{(i)} \sim \mathcal{N}(\mathbb{E}[Y^{(i)} \mid X], 1)$ with $i \in \{0, 1\}$. Performance of the estimates is evaluated by the means of two measures. The first measure is the mean-squared error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\tau(x_i) - \widehat{\tau}(x_i))^2. \tag{26}$$

The second measure is the coverage probability, which is defined as the probability of the estimated treatment effect $\widehat{\tau}(x)$ lies in the confidence interval of the true treatment effect $\tau(x)$. The target coverage is set to 0.95.

Experiment 1

In practice, there is often correlation between the treatment assignment and the potential outcomes, creating bias unless the methods used accurately adjust for this. To simulate this setting, the treatment effect is fixed at $\tau(x) = 0$ in the first experiment, the propensity score is set as

$$e(X) = \frac{1}{4}(1 + \beta_{2,4}(X_1), \tag{27}$$

where X_1 is the first feature and $\beta_{2,4}$ is the density function of the beta distribution with shape parameters 2 and 4. Finally we set the main effect as

$$m(X) = 2X_1 - 1. \tag{28}$$

Since the main goal of this experiment is to perform accurate propensity matching we use propensity trees (Procedure 2) as the base learner, growing $T = 1000$ trees with subsample size $s = 50$. We simulated $n = 500$ observations and varied d between 2 and 30. For the BART model we used the default settings for the depth prior, that is we set $\alpha = 0.95$ and $\beta = 2$. Furthermore, we set the number of trees $m = 50$ and $k = 3$. We also use the default hyperparameter choice for the σ prior, meaning we set $\nu = 3$ and $q = 0.90$.

Experiment 2

The second goal we want to achieve is to be able to identify neighbourhoods over which the actual treatment effect $\tau(x)$ is relatively stable. Therefore in the second experiment we evaluate the causal random forests performance when $\tau(x)$ is heterogeneous when holding $e(x) = 0.5$ and $m(x) = 0$ constant. Holding the propensity constant implies that the treatment assignment is independent of the features of any observation, reducing the simulation to a randomized experiment. We set $\tau(x)$ to be a smooth function on the first two features, X_1 and X_2, as

$$\tau(x) = f(X_1)f(X_2), \tag{29}$$

where

$$f(x) = 1 + \frac{1}{1 + \exp\left(-20(x - 1/3)\right)}. \tag{30}$$

Since the goal of this experiment is to assess performance of the causal forests, we use the causal random forest (Procedure 1) as base learner, growing $T = 2000$ trees with subsample size $s = 2500$. We simulated $n = 5000$ samples and varied d between 2 and 8. For the BART model, we use the same setting as experiment 1 however we now use 200 trees, that is $m = 200$.

However, k-NN and random forests are known to fill valleys and lower peaks of the true treatment effect function $\tau(x)$. This shortcoming is most apparent at the edges of the feature space. In order to demonstrate this, we perform a similar approach as described above, except now we set the function $f(x)$ as

$$f(x) = \frac{2}{1 + \exp\left(-12(x - 1/2)\right)}, \tag{31}$$

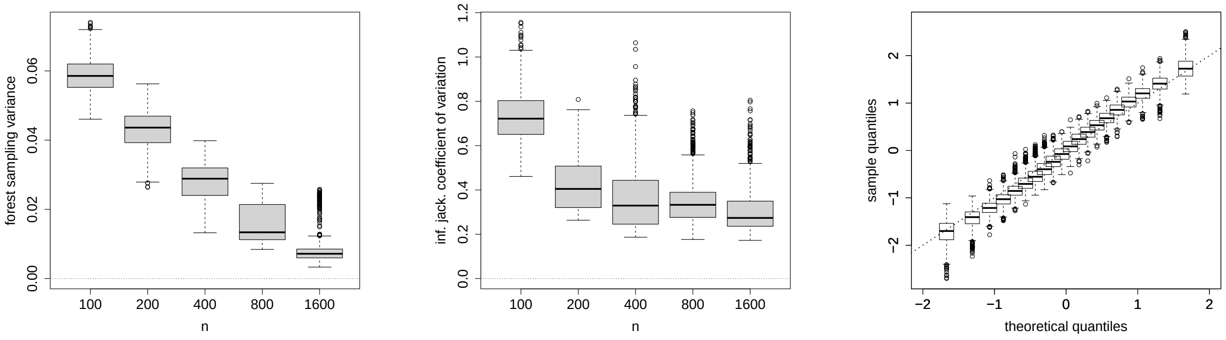
such that $\tau(x)$ has a sharper spike around the region where x_1 and x_2 are approximately 1. For this final experiment we keep the settings for the random forest the same. The settings of the BART model also remain the same except we now use 300 trees, that is $m = 300$.

4 Results

In this section, we will first provide results to support the theoretical findings. Second, we will display the outcomes of the three simulation experiments described in the previous section. Finally, a comparison of the models will be made on observational data.

4.1 Theoretical findings

The left panels, (a) and (b), of Figure 3 display graphical diagnostics for the behaviour of the sampling variance $\sigma_n^2(x)$ and the infinitesimal jackknife estimator, $\widehat{V}_{\text{IJ}}(x)$. Panel (a) shows that the sampling variance decreases to 0 when n increases, which suggests that Equation 8 holds, implying asymptotic normality. This is confirmed in panel (c) where the QQ-plot strongly resembles a perfect linear relationship. The middle panel shows that the RMSE of the estimator compared to the sampling variance, that is $\sqrt{\mathbb{E}\big[(\widehat{V}_{\mathrm{IJ}}(x) - \sigma_n^2(x))^2\big]}/\sigma_n^2(x)$, decreases for increasing n.



(a) Variance decay

(b) RMSE

(c) Quantile plot

Figure 3: Graphical diagnostics for asymptotic normality

These figures offer graphical diagnostics on the behaviour of the sample variance $\sigma_n^2(x)^2$ in the left panel, the middle panel displays the decline of the relative RMSE of the Inifinitesimal Jackkinfe compared to the sampling variance and the left panel shows a QQ-plot of sample quantiles, computed using Equation 8, versus theoretical standard normal quantiles. Panels (a) and (b) were computed over 50 simulation replications, the forest was trained on training samples with varying size as displayed on the horizontal axis and tested on 1,000 randomly drawn $d = 20$ values for x. Panel (c) was computed over 20 simulation replications, the forest was trained on $n = 800$ and tested on $n = 1000$ randomly drawn $d = 20$ values for x.

4.2 Simulation experiments

In experiment 1 we kept the treatment effect fixed at $\tau(x) = 0$ and allowed correlation between the treatment assignment and the propensity score. The results of this experiment are displayed in Table 1. First of all we note that BART has the lowest MSE in the lower dimensional cases. Both the random forest and BART outperform k-NN with $k = 10, 100$, implying they provide more unbiased estimates of the treatment effect. The coverage of the random forest is best in the low dimensional case of $d = 2$ and $d = 3$, reaching the target coverage of 0.95, however this is the only time it is reached as the coverage decreases when the dimensionality increases, the latter is also consistent for the nearest neighbour algorithms. BART convincingly reaches the target coverage independent of dimensionality.

The pattern described in Table 1 extends reasonably to the second experiment. The results of which are displayed in Table 2. The MSE of the random forest decreases when increasing the dimensionality for the lower dimensions, while still maintaining the 0.95 target coverage. This finding can be explained by the innate property of flexibility of the random forest. When dimensionality increases, individual trees have more flexibility in placing their splits to partition the data, this causes less similar trees in the forest, which leads to lower correlation between the trees. Furthermore, the

	MSE								Coverage							
d	**RF**		**BART**		**10-NN**		**100-NN**		**RF**		**BART**		**10-NN**		**100-NN**	
2	0.024	(1)	0.015	(1)	0.204	(2)	0.092	(2)	**0.947**	(5)	**0.991**	(2)	0.934	(2)	0.601	(11)
5	0.017	(1)	0.011	(1)	0.240	(3)	0.114	(3)	**0.947**	(5)	**0.993**	(2)	0.923	(2)	0.532	(11)
10	0.016	(1)	0.010	(1)	0.288	(3)	0.125	(3)	0.926	(8)	**0.992**	(3)	0.905	(2)	0.496	(12)
15	0.016	(1)	0.008	(1)	0.308	(3)	0.127	(3)	0.918	(8)	**0.996**	(2)	0.900	(2)	0.494	(12)
20	0.019	(1)	0.007	(1)	0.326	(4)	0.136	(3)	0.867	(11)	**0.997**	(1)	0.893	(2)	0.453	(12)
30	0.020	(1)	0.005	(0)	0.337	(4)	0.136	(3)	0.844	(11)	**0.999**	(1)	0.889	(2)	0.457	(12)
40	0.024	(1)	0.005	(0)	0.337	(4)	0.13	(3)	0.792	(14)	**0.999**	(1)	0.889	(2)	0.470	(12)

Table 1: Output for simulation experiment 1

This table reports the MSE and coverage of the random forest (RF), BART and k-NN with $k = 10, 100$ for experiment 1. The reported values are means calculated over 500 simulations repetitions. The integers between brackets multiplied by 10^{-3} denote the standard error of the simulation. **Bold** values indicate the target coverage is hit.

product of the variance of the individual trees multiplied by the correlation of all the trees equals the variance of the random forest. The increase in flexibility of individual trees therefore lowers the variance of the random forest which allows for a lower MSE. This property does not extend to BART, due to its additive nature, therefore we note that BART has a slightly increasing MSE. Furthermore, we find that the coverage decreases when dimensionality increases for the random forest and for k-NN. The coverage BART dropped compared to the first experiment but it still managed to hit the target coverage every time, whilst providing the lowest MSE estimates of all the models.

	MSE								Coverage							
d	**RF**		**BART**		**7-NN**		**50-NN**		**RF**		**BART**		**7-NN**		**50-NN**	
2	0.039	(1)	0.018	(1)	0.282	(3)	0.041	(1)	**0.968**	(3)	**0.964**	(7)	0.930	(2)	**0.947**	(4)
3	0.030	(1)	0.017	(1)	0.289	(3)	0.052	(1)	**0.958**	(3)	**0.967**	(8)	0.926	(2)	0.919	(4)
4	0.026	(1)	0.017	(1)	0.296	(3)	0.075	(2)	**0.953**	(5)	**0.971**	(7)	0.926	(2)	0.856	(5)
5	0.026	(1)	0.018	(1)	0.312	(4)	0.113	(2)	0.929	(6)	**0.970**	(6)	0.922	(2)	0.760	(5)
6	0.026	(1)	0.019	(1)	0.337	(3)	0.149	(2)	0.915	(9)	**0.956**	(9)	0.913	(2)	0.684	(6)
8	0.027	(1)	0.019	(1)	0.383	(4)	0.214	(3)	0.897	(9)	**0.967**	(6)	0.895	(2)	0.566	(5)

Table 2: Output for simulation experiment 2

This table reports the MSE and coverage of the random forest (RF), BART and k-NN with $k = 7, 50$ for experiment 2. The reported values are means calculated over 25 simulations repetitions. The integers between brackets multiplied by 10^{-3} denote the standard error of the simulation. **Bold** values indicate the target coverage is hit.

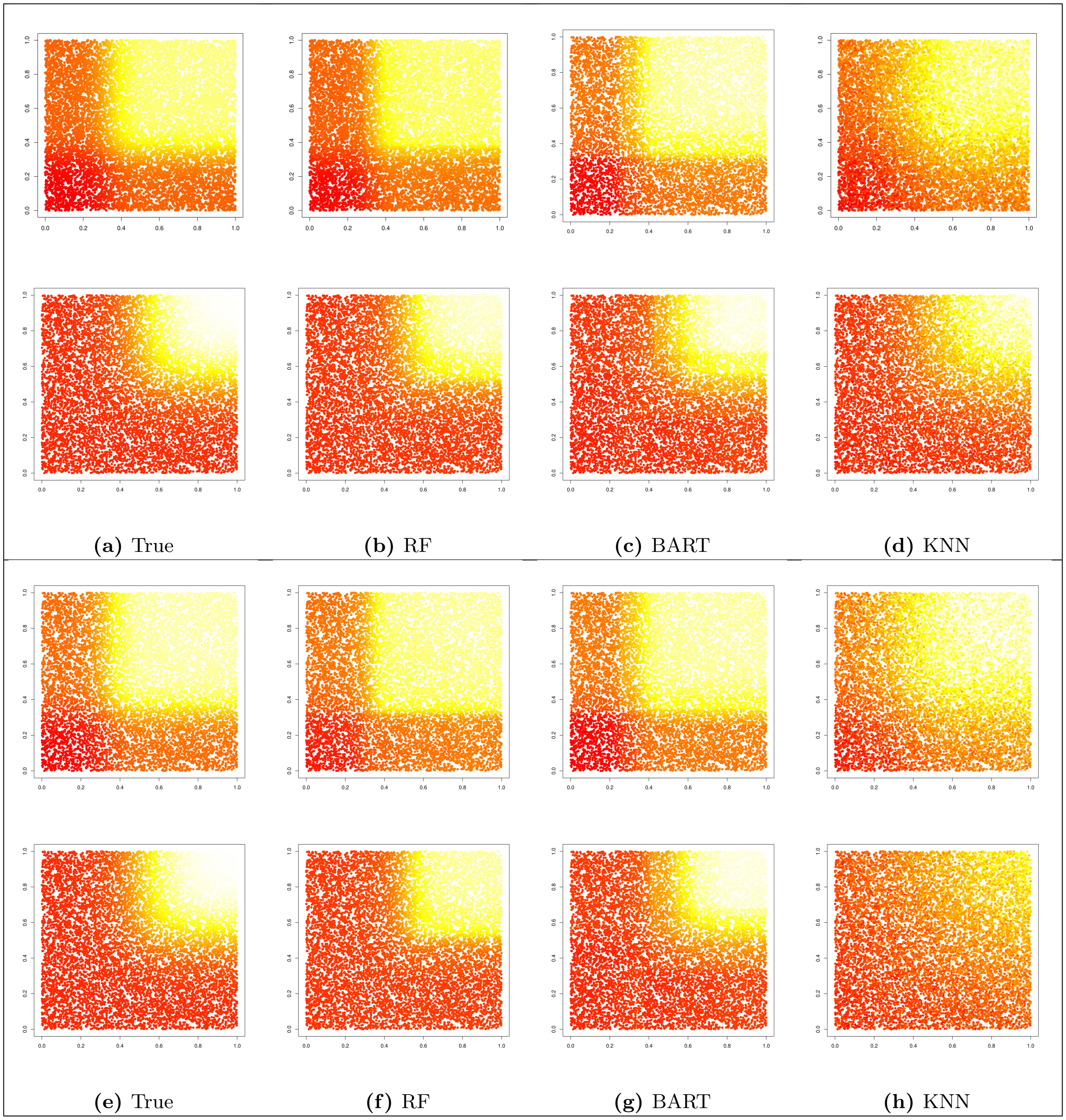
When comparing Table 2 with Table 3 we find some notable similarities and differences. Firstly, the MSE of the random forest decreases for lower dimensions, after The MSE is lower for a substantial amount of model predictions, when compared to Table 2. Since this is also apparent for the k-NN this could be caused by the experimental setup, $\tau(x)$ is easier to estimate for a large part of the sample. Secondly, the target coverage is not hit once by the random forest and decreases rapidly with d. This implies that the random forest is affected by bias.

	MSE				Coverage			
d	**RF**	**BART**	**10-NN**	**100-NN**	**RF**	**BART**	**10-NN**	**100-NN**
2	0.018 (0)	0.017 (0)	0.200 (2)	0.021 (0)	0.936 (3)	**0.953** (4)	0.935 (2)	**0.946** (2)
3	0.016 (1)	0.018 (1)	0.202 (2)	0.030 (1)	0.898 (6)	**0.945** (5)	0.934 (1)	0.898 (4)
4	0.018 (1)	0.018 (1)	0.209 (2)	0.054 (1)	0.857 (8)	**0.946** (6)	0.932 (1)	0.790 (4)
5	0.019 (1)	0.018 (1)	0.222 (2)	0.094 (1)	0.811 (10)	0.941 (6)	0.927 (2)	0.666 (5)
6	0.020 (1)	0.018 (1)	0.240 (2)	0.145 (2)	0.796 (9)	**0.944** (4)	0.921 (2)	0.579 (5)
8	0.026 (1)	0.019 (1)	0.291 (2)	0.263 (3)	0.716 (14)	0.937 (6)	0.900 (2)	0.447 (6)

Table 3: Output for simulation experiment 2, sharp $\tau(x)$

This table reports the MSE and coverage of the random forest (RF), BART and k-NN with $k = 10, 100$ for experiment 2, with the sharp $\tau(x)$. The reported values are means calculated over 40 simulations repetitions. The integers between brackets multiplied by 10^{-3} denote the standard error of the simulation. **Bold** values indicate the target coverage is hit.

To further understand this phenomenon we plotted the response surfaces in Figure 4. When comparing the rows in the top window we find that the random forest and BART clearly capture the true response surface better than k-NN. However the second row shows that the random forest is having trouble to accurately capture the peak in the top right of surface, whereas BART captures this more accurately. We see that this pattern extends to the bottom pannel, which displays a higher dimensional problem. In this case, they perform similar to the low dimensionality while the performance of k-NN decreases more. This illustrates that BART is noticable less affected by the bias at the edge of the feature space. Therefore, it is able to maintain the target coverage rate for more than half of the tested dimensions. However, we do note that the coverage tends to be decreasing, implying that BART would get affected by bias in higher dimensional problems.



(a) True (b) RF (c) BART (d) KNN

(e) True (f) RF (g) BART (h) KNN

Figure 4: Graphical representation of the treatment effect function response surface

The top half, with columns (a) through (d), corresponds to a low dimensional setting with $d = 6$. The bottom half, with columns (e) through (h), corresponds to a high dimensional setting with $d = 20$.The first and the second row represent the response surfaces for function $\tau(x)$ corresponding with Table 2 and Table 3 respectively. The colors indicate values of $\tau(x)$ for the first column and $\widehat{\tau}(x)$ for the other columns. On the scale, a red color implies a low value where a (white) yellow color implies a high value.

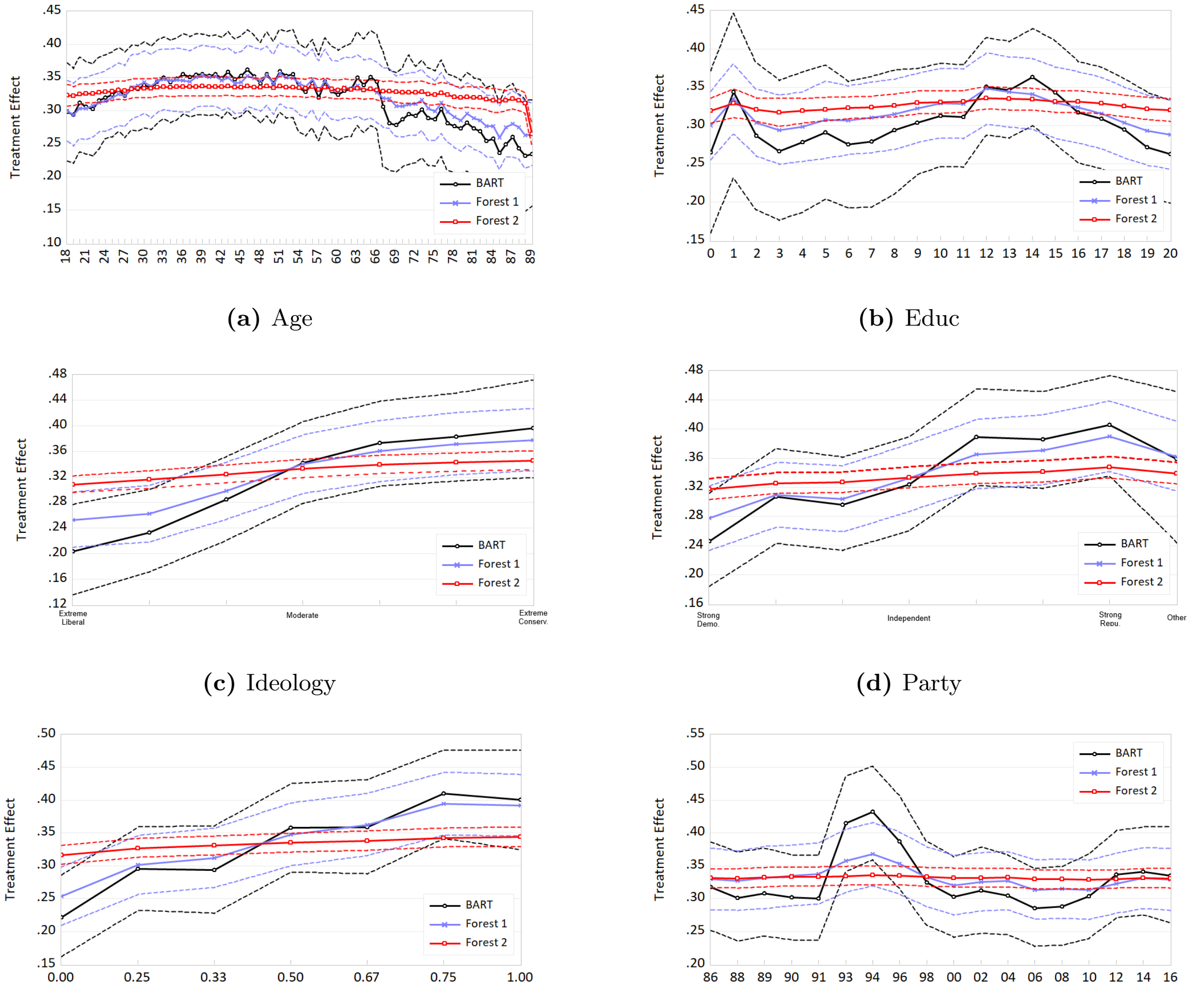
4.3 Model comparison based on observational data

For our BART model we used the default settings for the depth prior, that is we set $\alpha = 0.95$ and $\beta = 2$. Furthermore, we set the number of trees $m = 100$ and $k = 3$. We also use the default hyperparameter choice for the σ prior, meaning we set $\nu = 3$ and $q = 0.90$. We will compare the behaviour of the random forest to BART. Wager and Athey (2018) proposed two procedure to construct random trees, the first procedure is the double-sample tree and the second procedure is the propensity tree, recall that when taking an ensemble of these trees our random forest model is formed. We will compare differences between these two procedure as well as to BART. For both procedure we set the number of trees $T = 2000$. For the double-sample forest we set the subsamplesize $s = 0.05n$ and for the propensity forest we set the subsamplesize $s = 0.1n$, recall that $n = 24583$, which makes the subsamplezie $s = 1229$ and $s = 2458$ respectively.

With these settings we estimate the treatment effects for the individuals in the dataset. Since we are only interested in the estimation of the treatment effect and not in, for example out of sample performance, we first estimate the models ("train") for the entire dataset. Then we let the models predict ("test") on the same dataset, this will return the models estimates for the treatment effect since the structure of the models are locked in this stage. The results are displayed in Figure 5, where each panel presents the treatment effect estimates when conditioned on one of the variables. Note that the models solely provide point estimates, the lines drawn in the graph are merely present to more clearly display the trend in the conditional treatment effect.

Firstly, the estimates provided by the propensity forest seem to be quite insensitive to changes in the variables. It only appears to slightly capture the trend, as can be seen by the curvature in panel (a) of Figure 5. The propensity forest has a very narrow distribution around its mean, this is displayed more clearly in Figure 6 panel (c), which displays the distribution of estimated treatment effect on individual level. 5 also shows that BART and the double-sample forest provide very similar estimations, where BART displays more sensitivity as it shows more sudden fluctuation than the double-sample forest. This is paired with more uncertainty about the point estimates for BART since the 95% confidence interval is broader than the confidence interval of the double-sample forest.

Given this information we can discuss model performance. Firstly, it is clear to see that the propensity forest (Forest 2) does not seem to accurately capture patterns in the data. Therefore, this model will be deemed as worst performing. The double-sample forest and BART provide very similar estimations. An argument could be made that since the double-sample forest has a smaller confidence interval, it provides the best point estimates. However, we have seen in the simulation experiment that the coverage probability of the forest is not guaranteed to be 0.95. Especially not in the case of a sharp peak around the edge of the feature space. Since it is unclear whether that is the case in this observational setting we find BART to provide the most reliable estimates. Therefore, we rank BART as the best performing model, followed by the double-sample forest.



(a) Age

(b) Educ

(c) Ideology

(d) Party

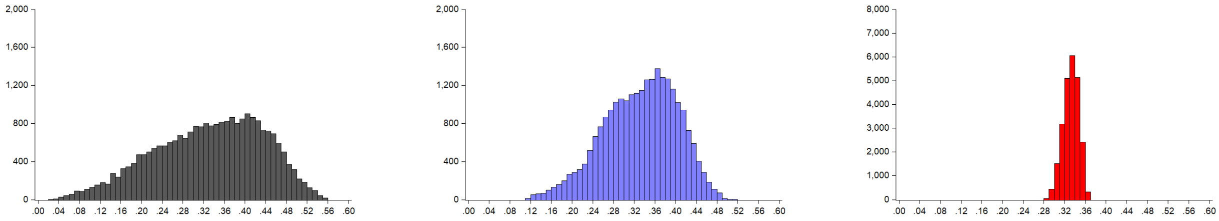
(e) Racist

(f) Year

Figure 5: Conditional treatment effect estimates $\widehat{\tau}(x)$

The solid lines represent mean individual treatment effect estimates aggregated over the characteristic displayed on the horizontal axis. So the first observation in the top left panel corresponds with the mean treatment effect of respondents with age 18. The black, blue and red line correspond with the models **BART**, random forest computed according to procedure 1 (**double-sample forest**) and random forest computed according to procedure 2 (**propensity forest**) respectively. The dotted lines represents the 95% confidence interval for the estimates.

For the interpretation of Figure 5, recall that the individuals in the control group and treated group received a question about money spent on "assistance to the poor" and "welfare" respectively. Since all the conditional treatment effect estimates, displayed in the graphs, are positive we can conclude that the average treatment effect is also positive, which takes a value of around 0.33 for all the models. This means that on average the probability that an individual thinks too much money is spent on welfare is higher than the probability that an individual thinks too much money is spent on assistance to the poor. When conditioning on the characteristics of individuals we find some notable patterns. Especially for the characteristics political ideology, party affiliation and racist tendencies, displayed in panels (c), (d) and (e) respectively. We can clearly see an up-sloping trend for the treatment effect for these characteristics. This means that individuals that are more conservative have a higher treatment effect compared to more liberal individuals. The same holds for party affiliation. Note that in the American political system these two variables are often closely related, liberal often corresponds with democratic affiliation and conservative often corresponds with republican affiliation. Individuals with more racist tendencies have a higher treatment effect than individuals with less racist tendencies. For the other characteristics, age of the individual, years of education and year of the survey, there is no distinct pattern visible. However, there are increased treatment effect estimates for the years 1993, 1994 and 1996, shown in panel (f). This peak may be caused by President Bill Clinton's first term in office (1993-1996). Clinton actively advocated for welfare reform policies. Therefore, the difference between assistance to the poor and welfare might have been increased due to the presence of the term welfare in Clinton's campaign and during his presidency. This in turn causes increased treatment effect estimates.



(a) BART

(b) Double-sample forest

(c) Propensity forest

Figure 6: Estimated treatment effect probability distribution

The frequency is displayed on the vertical axis, the treatment effect estimate is displayed on the horizontal axis in bins with width of 0.01. The colours of the histogram correspond with the colours used in 5. Note that the scale of panel (c) differs with a factor 4.

5 Conclusion

Focusing on the issue of treatment effect estimation we have compared the performance of multiple methods in both a simulation experiment and an observational study. We have introduced three methods: a double-sample random forest, a propensity forest and BART, to answer the research question:

RQ: *How do causal random forests perform compared to Bayesian Additive Regression Trees (BART) when exposed to observational data?*

We have found that, based on the simulation experiment, the random forests are outperformed by BART in both mean-squared error (MSE) and statistical coverage. However, the random forests do, in general, outperform a baseline k-nearest neighbour algorithm. One of the main flaws of the random forest is that when the response surface of the treatment effect gets sharp around the edges of the feature space, the performance of the random forest tends to diminish rapidly.

When comparing the performance of the three methods when exposed to observational data, we find that the double-sample forest and BART provide similar, realistic, estimations. The propensity forest does not seem to manage to capture the patterns present in the data. Even though the double-sample forest and BART provide similar estimates, we would still suggest the use of BART in practice. Since, it has shown to consistently hit the target coverage rate in the simulation experiment and to maintain the lowest MSE in general, even when the response surface of the treatment effect gets sharp around the edges of the feature space.

5.1 Future research

It is known that in machine learning the choice of hyperparameters can have substantial effect on the performance of the methods. Due to computational constraints we did not include any hyperparameter tuning in our research and used default values suggested by Chipman et al. (2010), Green and Kern (2012) and Wager and Athey (2018). Hence, in future research, hyperparameter tuning could be performed to evaluate whether this positively impacts the performance of any of the methods used. Secondly, we compared the random forests to one competing model. For further research, a broader array of models could be compared on multiple varying simulation experiments and multiple observational studies. This creates an overview of the treatment effect estimation repertoire, which allows researchers to select models that fit their research setting best.

References

S. F. Assmann, S. J. Pocock, L. E. Enos, and L. E. Kasten. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069, 2000.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

H. A. Chipman, E. I. George, R. E. McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

D. P. Green and H. L. Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

L. A. Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics*, pages 1449–1458, 1972.

J. Pearl. Causal inference. *Causality: Objectives and Assessment*, pages 39–58, 2010.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Appendix A

Below is the formulation of the questions asked on the survey.

- *Main question*: We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount. First, are we spending too much, too little, or about the right amount on **assistance to the poor** (**welfare**)?

- *Ideology*: We hear a lot of talk these days about liberals and conservatives. I'm going to show you a seven-point scale on which the political views that people might hold are arranged from extremely liberal–point 1–to extremely conservative–point 7. Where would you place yourself on this scale?

- *Party*: Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?
  - Strong democrat (0)
  - Not strong democrat (1)
  - Independent near democrat (2)
  - Independent (3)
  - Independent near republican (4)
  - Not strong republican (5)
  - Strong republican (6)
  - Other party (7)

- *Racist*: On the average (Negroes/Blacks/African-Americans) have worse jobs, income, and housing than white people.
  - Do you think these differences are mainly due to discrimination? ("yes"=0, "no"=1)
  - Do you think these differences are because most (Negroes/Blacks/African-Americans) have less in-born ability to learn? ("yes"=1, "no"=0)
  - Do you think these differences are because most (Negroes/Blacks/African-Americans) don't have the chance for education that it takes to rise out of poverty? ("yes"=0, "no"=1)

- Do you think these differences are because most (Negroes/Blacks/African-Americans) just don't have the motivation or will power to pull themselves up out of poverty? ("yes"=1, "no"=0)