Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

SIMULATION STUDY IN QUANTIFYING HETEROGENEOUS CAUSAL EFFECTS

Jianing Zhao

Department of Computer Science
College of William and Mary
251 Jamestown Rd
Williamsburg, VA 23185, USA

Daniel M. Runfola

AidData
427 Scotland St
Williamsburg, VA 23185, USA

Peter Kemper

Department of Computer Science
College of William and Mary
251 Jamestown Rd
Williamsburg, VA 23185, USA

ABSTRACT

Quantifying the impact of an intervention or treatment in a real setting is a common and challenging problem. For example, we would like to calculate the environmental implications of aid projects in third world countries that target economic development. For causal inference problems of this kind, the Rubin causal model is one of several popular theoretical frameworks that comes with a set of algorithmic methods to quantify treatment effects. However, for a given data set, we neither know the ground truth nor can we easily increase the size of the data set. So, simulation is a natural choice to evaluate the applicability of a set of methods for a particular problem. In this paper, we report findings of a simulation study with four causal inference approaches, namely two single tree approaches (transformed outcome tree, causal tree), and two random forest versions of the former.

1 INTRODUCTION

We frequently seek to test the effectiveness of targeted interventions - for example, a new website design, a medical treatment, or a third world aid project. This is important for informed policy decisions to allocate resources in a meaningful way.

The work presented here is based on the Rubin Causal Model (or potential outcome framework), where causal effects are estimated through comparisons between observed outcomes and the "counterfactual" outcomes one would have observed under the absence of an intervention (Imbens and Rubin 2015). In the causal inference literature, the terminology to describe this follows a medical point of view. The intervention is called a treatment and all observed units are separated into two subsets: a group of treated cases versus a group of control cases. The Rubin causal model is a common framework to model and evaluate causal inference. The outcome of such an analysis is either the average treatment effect observed on the whole population of units or the conditional average treatment effect one observes for a specific subset of units that share some particular characteristics. This leads to the need to describe units by relevant properties, i.e., values for a number of variables or features that are called covariates in this context. The interest in the conditional average treatment effect naturally arises from the fact that treatments are not prescribed

randomly or in general but to address an observed condition or situation. For example, a third world aid project to fund a hospital in general is good but for sure it will have a bigger impact - a bigger conditional average treatment effect - if placed in an area that is currently underserved with medical treatment facilities and where people frequently suffer from diseases that are easily curable. In observational studies, especially in the medical and social sciences, there is interest in the estimation of such heterogeneous causal effects.

If one leaves the realm of simulation studies, where one can generate experimental data for targeted treatment and control groups, and has to rely on real-world, observational data, one runs into a fundamental missing data problem. For any unit, we can only observe the unit with the treatment, or without the treatment, but not both at the same time. So, the ground truth for a causal effect can not be observed for any individual unit and its calculation is not directly possible in the Rubin causal model as the causal effect is the difference between the outcome for the treatment and control case. Several techniques have been developed to work around this fundamental crux in the Rubin causal model. For the conditional average treatment effect they all essentially compute differences between groups that are "similar" or are made "comparable" by some appropriate rescaling.

Many approaches to estimating heterogeneous effects have emerged over the last decade. LASSO (Tibshirani 1994) and support vector machines (SVM) (Vapnik 1998) may serve as two popular examples, however, we have limited this analysis to a small subset of techniques that are based on regression trees. Specifically, we test Transfomed Outcome Trees (TOTs), Causal Trees (CTs), Random Forest TOTs (RFTOTs), and Causal Forests (CFs). We follow the work of (Athey and Imbens 2015), who demonstrated how decision trees and random forests can be adjusted to estimate heterogeneous causal effects. While traditional tree-based approaches rely on training with data with known outcomes, Athey and Imbens illustrated that one can estimate the conditional average treatment effect on a subset with regression trees after an appropriate data transformation process. The historic lineage of causal inferential study using trees is relatively young, but rapidly growing. In (Su, Tsai, Wang, Nickerson, and Li 2009), Su et al. proposed a statistical test as the criterion for node splitting. In (Athey and Imbens 2015), Athey and Imbens derived TOTs and CTs, an idea that is followed up on by Wagner and Athey (Wager and Athey 2015) with CF (causal forest, random forests of CTs), and similarly Denil et al. in (Denil, Matheson, and de Freitas 2014) who use different data for the structure of the tree and the estimated value within each node. Random forests naturally gave rise to the question of confidence intervals for the estimates they deliver. Following this, Meinshausen introduced quantile regression forests in (Meinshausen 2006) to estimate a distribution of results, and Wagner et al (Wager, Hastie, and Efron 2014) provided guidance for confidence intervals with random forests. Several authors, including Biau (Biau 2012), recognize a gap between theoretical underpinnings and the practical applications of random forests.

Our ultimate goal is to apply these techniques to analyze a large data set for world bank aid projects that ranges over a time period of 30 years and covers locations worldwide. The research question is to estimate the impact on vegetation of third world aid projects that primarily aim at economic development. The data set is challenging to analyze for various reasons. Because we are unable to produce "ground truth" values to understand the accuracy of our approaches in this real-world case, here we turn to a simulation study on tree-based causal inference techniques. The key benefit of a simulation study is that we can design a stochastic model in such a way that we can generate data for the treated and control group as much as needed and we know the ground truth of the causal effect. The questions we want to answer in this way are: a) if a causal inference technique gives us a close estimate of the causal effect for a simulated data set similar in kind to the one we want to analyze, b) if we increase the number of covariates that impact the causal effect, how does this affect the accuracy of causal inference techniques c) if we increase the amount of available data, how quickly does the estimated causal effects converges to the ground truth, d) if we vary the required minimum number of control and treated units to compare for the calculation of the causal effect in the tree generation algorithms, how does this affect the accuracy of results

The rest of the paper is structured as follows. We introduce the four tree-based causal inference techniques in Section 2 followed by a description of our stochastic model to generate data in Section 3. In Section 4, we present the results of our simulation study and discuss our findings.

2 CAUSAL INFERENCE TECHNIQUES

Before we go through the details of tree-based causal inference techniques, we briefly introduce some notation for the Rubin causal model and recall its main concepts.

2.1 The Rubin Causal Model And Conditional Average Treatment Effects

Suppose we have a data set with n independently and identically distributed (iid) units with $i = 1, \cdots, n$. Each unit has an observed feature vector $X_i \in [0,1]^d$, with d covariates and a response (i.e., the outcome of interest) $Y_i \in \mathbb{R}$. A treatment is considered binary and is formalized with an indicator variable $W_i \in \{0,1\}$ for each unit i. For a unit-level causal effect, the Rubin causal model defines the treatment effect on unit i as $\tau_i = Y_i(1) - Y_i(0)$, the difference between treated $Y_i(1)$ and untreated $Y_i(0)$ outcome.

In this paper, we are interested in calculating the heterogeneous causal effect, which we define as $\tau(x) = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i = x\big]$ following (Hirano, Imbens, and Ridder 2003). In an observational study, a unit is either treated or not, so we know either $Y_i(1)$ or $Y_i(0)$, but not both. However, one can still estimate $\tau(x)$ if one assumes unconfoundedness: $W_i \perp (Y_i(1), Y_i(0)) \mid X_i$.

Unconfoundedness means that given some features X_i, the probability of outcomes $(Y_i(1), Y_i(0))$ is independent of the assignment of a treatment W_i. Under the unconfoundedness assumption, (Athey and Imbens 2015) show that one can calculate the causal effect as $\tau(x) = \mathbb{E}\big[Y^* \mid X_i = x\big]$, where the transformed outcome Y^* is defined as

$$Y_i^* = Y_i \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))}, \tag{1}$$

and the propensity score function $e(x)$ is defined as $e(x) = \mathbb{E}\big[W_i \mid X_i = x\big]$. In laboratory experiments as well as in simulation studies, it is common to randomly assign a treatment to a unit and to use $e(x) = 0.5$ to obtain balanced group sizes for treated and control groups. In reality, a certain disposition X_i will either qualify/demand for treatment such that a treatment will be assigned in most cases or disqualify for treatment and a treatment will be rarely assigned. The propensity score accounts for this effect. Several approaches to estimate the propensity score can be selected (Rosenbaum and Rubin 1983), (Ho, Imai, King, and Stuart 2007); for the world bank data set, for instance, we use logistic regression in order to provide a stronger comparison to econometric modeling approaches most commonly employed by the international development community today.

2.2 Regression Tree For Causal Inference

A regression tree is a binary tree to represent a step function $f : \mathbb{R}^d \to S$ with S being a finite subset of $\mathbb{R}$. Each of its leaf nodes carries a real value. Each of its internal nodes have an associated variable x (a covariate in our case) and a threshold t such that the edge to its left child carries a condition $x \leq t$ and its right child a corresponding condition $x > t$. So, a path from the root node to some leaf node encodes a conjunction of conditions along its edges, such that for any $x \in \mathbb{R}^d$ that satisfies all conditions along that path, the value at the leaf node gives $f(x)$.

An algorithm to compute a regression tree such as CART takes a set of sample tuples $(x, f(x))$ (the training data) and creates a tree by starting at a root node and recursively splits nodes by identifying a variable and a threshold to add left and right children to a node. The key step is then to have a rule to decide if a current leaf node should become an internal node by adding left and right children and how to determine the variable and threshold for this node. A good splitting rule partitions data from the parent node into left and right child nodes so that the resulting homogeneity of the child nodes is an improvement over the parent node. The splitting rule typically follows a greedy strategy and selects the covariate and

threshold that gives the greatest improvement. The rule is also complemented by a stopping criteria to avoid arbitrary fine partitions and very large trees.

Note that a typical outcome is a tree that does not consider all d covariates on a path from root node to a leaf node. This implies that the algorithm selects only a subset of covariates that matter for the resulting $f(x)$. This leads to a notion of relevance for covariates.

One can also look at a tree as a way to partition a data set in a number of bins such that each leaf node has an associated bin of data points whose covariate values satisfy the conjunction of conditions along the path from the root node to that leaf node. This leads to the understanding that data points in the same bin are considered "similar" or "comparable".

In order to adopt the concept of a regression tree and its algorithms, one need to adjust the splitting rule in a way that data points in the bin of a leaf node either have the same causal effect or they can be used to compute a conditional average treatment effect for all elements in that bin.

2.3 Transformed Outcome Tree

A TOT is a regression tree that uses Y^* for $f(x)$. As mentioned above, the transformed outcome is calculated with (1), then a traditional regression tree method is employed to generate a TOT for the given data set. The causal effect is subsequently estimated with $\tau(x) = \mathbb{E}\left[Y^* \mid X_i = x\right]$, where feature vector x leads to a leaf of the tree and $\hat{\tau}(x)$ is the average transformed outcome of all units in the corresponding leaf of the tree.

2.4 Causal Tree

When a regression tree is used to estimate heterogeneous causal effects, tree construction is a key step. In a classic regression tree, mean square error (MSE) is often used as the criterion for node splitting, and the average value within the node is used as the estimator. Following Athey and Imbens (Athey and Imbens 2015), we use (2) as the estimator, and we replace the traditional MSE by summing $Y_i^* - \hat{\tau}(X_i)$.

$$\hat{\tau}(X_i) = \sum_{i \in T} Y_i \cdot \frac{W_i / \hat{e}(X_i)}{\sum_{j \in T} W_j / \hat{e}(X_j)} - \sum_{i \in C} Y_i \cdot \frac{(1 - W_i)/(1 - \hat{e}(X_i))}{\sum_{j \in C}(1 - W_j)/(1 - \hat{e}(X_j))}, \tag{2}$$

where T represents treatment units, and C control units. This new error term is then used to split the tree in a way identical to traditional regression trees.

Due to characteristics of the data in our applied study, we found that cases where only C or T units existed in children naturally emerged, which could lead to inaccurate estimates. Specifically, using the above split criterion, we were not able to estimate the causal effect as no counterfactual cases (or treated cases) exist. Following this, we introduce an additional constraint on the tree which prevents node splitting when the resultant child will result in all T or C cases.

2.5 Random Forest

A random forest is a set of regression trees that are independently built using a bootstrap sample of the given data set and the splitting criterion applied in the construction only selects the best split among a subset of all possible predictors randomly chosen at that node (Breiman 2001). The random forest concept can be applied to any approach that computes a single tree, as it is the case for the TOTs or CTs. To obtain an estimate for $\hat{\tau}(x)$ from a forest of trees, one calculates the average of all individual estimates $\hat{\tau}(x)$ that one obtains from each single tree in the forest. The key idea is that the trees represent the variability in the construction of trees due to the variability in the data (exposed by the bootstrapping and random selection of a subset of candidate covariates for a split) and that the averaging for the overall result creates a robust estimate for $\hat{\tau}(x)$ thanks to the large number of independently generated trees.

We rely on the randomForest R package that implements Breiman's approach and implemented a causal forest algorithm with the help of the scikit learn package.

3 SIMULATION MODEL

We want to iteratively simulate synthetic datasets with known parameters to evaluate how the number of covariates, the dataset size, and other parameters impact the accuracy of results of causal inference techniques. In order to do so, we need a model that first and foremost gives us the ground truth about the causal effect. In addition to that, we need a number of d covariates that contribute to an outcome Y_i for each of the data points $i = 1, ..., n$ that we need to generate from that model. To do this, we assume an additive model and follow a bi-partite data generation process, in which two equations are used (one for treated cases and another for control cases). For the treated cases, we use

$$Y_i(1) = W_i * (c + \sum_{i=1}^{k} \beta_i x_i) + \sum_{i=k+1}^{d} \beta_i x_i + \beta_0 + \varepsilon = c + \sum_{i=1}^{d} \beta_i x_i + \beta_0 + \varepsilon, \tag{3}$$

with treatment indicator variable W_i set to 1. We have a set of covariates $x_1, ..., x_k$ that contribute to the treatment effect in addition to the first term c that creates an constant effect regardless of any covariate setting. The equation includes further $d-k$ covariates and an error term ε. The $d-k$ covariates act as a distractor to the identification of variables that are relevant for the treatment effect. Constant c and variable ε are not observable and not represented in the data given to the causal inference calculation. Constant $c > 0$ provides some positive treatment effect regardless of covariates, which avoids diminishing treatment effects if covariate values are close to zero.

For the control cases, we use

$$Y_i(0) = W_i * (c + \sum_{i=1}^{k} \beta_i x_i) + \sum_{i=k+1}^{d} \beta_i x_i + \beta_0 + \varepsilon = \sum_{i=k+1}^{d} \beta_i x_i + \beta_0 + \varepsilon, \tag{4}$$

with treatment indicator variable W_i set to 0, such that the first k covariates do not impact the outcome.

So, the ground truth for each unit i is $\tau_i = Y_i(1) - Y_i(0) = c + \sum_{i=1}^{k} \beta_i x_i$. This allows us to observe heterogeneity in the causal effect as $\tau(x)$ depends on the value settings of $x_1, ..., x_k$. Since $\tau(x)$ does not depend on value settings for $x_{k+1}, ..., x_d$, we can also see if a causal inference result is consistent with this.

To generate a data set with some randomly assigned treatment and a given propensity function $e(x)$, we sample a random vector $(x_1, \ldots, x_d, \varepsilon)$ and with probability $e(x)$ select equation $Y_i(1)$ or $Y_i(0)$ otherwise to compute the outcome value for the data point. We sample values for each of the d covariates and ε independently from a probability distribution and its parameter settings associated with the particular covariate or error term. So for $e(x) = 0.5$, we use each of the two equations to produce about one half of all data points. $Y_i(1)$ gives the result for treated cases; $Y_i(0)$ is for the control group.

The use of different distributions across covariates implies that covariates are not operating at the same scale. We want to use realistic values for the synthetic data set but for the calculation of Y we want covariates contribute in a similar manner. To standardize each covariate with feature scaling $(X - min(X_i))/(max(X_i) - min(X_i))$, we define $\beta_i = 1/(max(X_i) - min(X_i))$ and $\beta_0 = -\sum_{i=k+1}^{d} min(X_i)/(max(X_i) - min(X_i))$. For the treatment effect, $c = c' - \sum_{i=1}^{k} min(X_i)/(max(X_i) - min(X_i))$, such that we still have some extra c' to see a constant treatment effect. Regression trees themselves are not sensitive to scaling, so we need not scale our synthetic data set for the causal inference techniques. Sampling from a non-uniform distribution implies that some values will occur more frequently than others, which will create a data set that provides ample set of samples for some cases and but only few for others. This is expected to be the case for real data sets, which is why we want to see how a causal inference technique reacts to this.

Our model meets the assumptions that we made for causal inference: the data points are i.i.d and unconfoundedness is fulfilled by a random treatment assignment given x. In addition to the ability to create realistic synthetic data of arbitrary size, we are also in the position to support an artificial best case scenario where the data set contains pairs of treated and untreated units to test the causal inference techniques.

3.1 Configuring The Simulation Model To Approximate The Real Data Set

The real data set that we ultimately want to analyze is on World Bank aid projects and the real challenge is to quantify the impact of these projects on the environment. The data is based on data of World Bank projects with covariates describing the project's amount of funding, its beginning and duration. In order to analyze the impact of World Bank projects on the environment, the data set has been enhanced with information on the geographic location of World Bank projects and satellite derived information on the geographic, environmental, and economic characteristics of each project location. So covariates include longitude and latitude for location, elevation and slope to describe the terrain, distance to rivers and roads as well as population density for economic characteristics, annual minimum, maximum and average values for air temperature and precipitation for the last 30 years and finally an index for vegetation cover (Normalized Difference Vegetation Index, NDVI). A linear regression has been used to aggregate time series data into covariates for intercept and slope.

The existing data set has d=37 covariates and 16369 data points and the challenge is to recognize the difference in average NDVI values before and after a project starts as a treatment effect. Of course, the ground truth for the real data is not known. For the simulation study, we want to see if tree-based causal inference techniques apply to a data set of this kind. We choose d=37 for the simulation model and for each covariate in our data set, we perform a distribution fitting. Figure 1 shows an example for the variable that describes the average of annual maxima of NDVI values for all years before the start of the project with a q-q plot on the left and that shows to what extent the observations match with a Normal distribution. The right side of Figure 1 shows a heat map of the correlation among all covariates and correlations are moderate as the predominantly cooler colors indicate. The individual covariates show limited correlation such that we can sample individual entries for $(x_1,\ldots,x_d,\varepsilon)$ independently and from a distribution that we fitted for each covariate. For the error term ε we decided to use a normal distribution $\mathcal{N}(0,1)$. For the treatment effect in the simulation model, we need to choose a value for k and select the covariates for $\tau_i = c + \sum_{i=1}^{k} x_i$. We follow the common 80/20 rule that suggests that the vast majority of an effect (about 80%) is caused by a small minority of parameters (about 20%). So, we consider different scenarios with $k = 1$, 2, 4, and 8 variables and select a random subset k covariates out of $d = 37$ for our experiment. In this way, we have a stochastic input model that we can use to derive data sets of any size with a known ground truth on treatment effect as well as the role of individual covariates that are responsible for the effect.

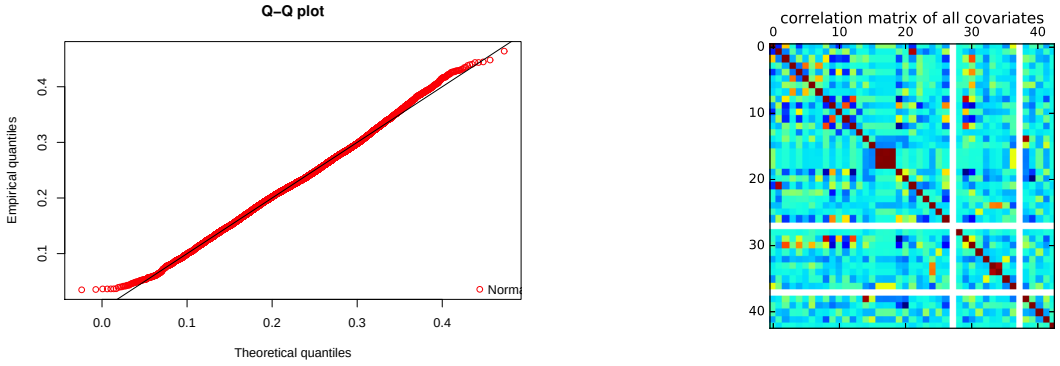


Figure 1: A q-q plot for fitting a distribution to the annual averages of maximum NDVI values before the start of a project (left), heat map of correlation among all covariates in the data set (right).

4 EXPERIMENTS AND RESULTS

In this section, we study the ability of CT, CF, TOT and RFTOT approaches to calculate heterogeneous treatment effects for our simulated data set. All of these approaches are based on variants of regression trees, which essentially are able to approximate a continuous function $f : R^d \rightarrow R$ with a step function based on some interval partitioning of the domain of f. The stochastic model to generate the data has a treatment effect function $\tau(x) = c + \sum_{i=1}^{k} \beta_i x_i$, which at least in principle allows for a discretized approximation with a step function in a straightforward manner as the ranges of values for each x_i can be partitioned into intervals and each interval contributes its average to the overall outcome. We begin our evaluation with a basic comparison of the accuracy of its approximation of $\tau(x)$ that each approach achieves for a given data set.

4.1 Accuracy Of Causal Effect Estimate

Before we look into a realistic data set, we want to report on reassuring findings for a more common and simpler model that is closer to the ones typically analyzed in the literature. We exclusively use standard normal distributions $\mathcal{N}(0,1)$ for all covariates and ε which implies $\beta_i = 1$ for $i > 0$ and $\beta_0 = 0$ for Eqs. 3 and 4. We analyze a model configuration for $n = 2000$, $k = 1$, $d = 9$, and $c' = 1$. We use a random treatment assignment such that each unit has the same probability to be treated, $e(x) = 0.5$. On the left, Figure 2 shows box plots for the treatment effect $\tau(x)$ in column TRUTH as well as estimates $\hat{\tau}(x)$ for CF, CT, TOT and RFTOT in corresponding columns. The distribution of $\tau(x)$ shows that the treatment effect is heterogeneous. The resultant distributions all encompass the true mean results, but with considerable difference in overall metrics of error. In line with published results, the Causal Forest approach is the most accurate across all simulations, random forest variants perform better than single tree variants and the CT is better than the TOT. Of course, the accuracy is expected to depend on the amount of data that is available. So, we also test the convergence of each method as the size of data increases, as shown in the middle diagram in Figure 2. We measure the mean squared error (MSE) between estimated $\hat{\tau}(x)$ and the ground truth of $\tau(x) = c + \sum_{i=1}^{k} \beta_i x_i$ to evaluate the accuracy of the estimates. It shows the MSE of each method with increasing data size, while the diagram to the right shows a zoomed-in version of the MSE of the CF approach (due to the lower magnitude of MSE observed). As expected, accuracy improves with the amount of data, random forest approaches generally show a better performance than single tree approaches and CF shows best results. The box plots in the right diagram of Figure 2 result from repeated calculations and show the variability of MSE in response to the variability in inputs as well as the general convergence for increasing values of n.

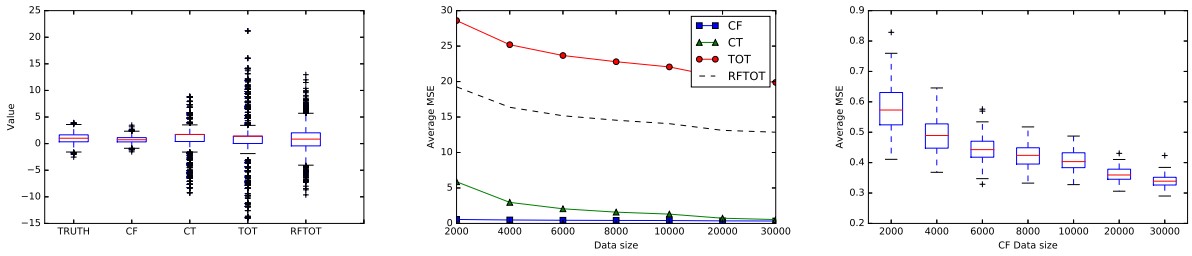
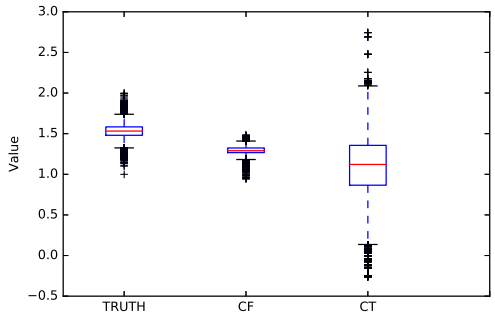


Figure 2: Estimated treatment effects (left), MSE changes with data size (middle), and for CF (right).

Seeing positive results that are consistent with original research publications motivates us to move forward and analyze scenarios closer to our real data set. The first scenario we examine considers a simulated dataset with a randomized treatment assignment (each unit has the same probability to be treated, $e(x) = 0.5$). For our stochastic model, we used $k = 1$, $d = 37$, $c' = 1$, and $n = 20000$. As we increase the number of covariates d, we report findings for two variants, one still samples values from a normal

distribution $\mathcal{N}(0,1)$ independently for each covariate and a second one where all covariates are sampled independently from the distributions that we fitted to the real data set. So the second configuration models the situation where a single covariate is responsible for the treatment effect in a setting that resembles our real data set in terms of dimensions ($d = 37$), covariate distributions, and size ($n = 20000$). Note that $E[\tau(x)] = 1$ for the selected parameter settings and sampling distributions for covariates, in particular due to the settings of β_i coefficients and $c' = 1$. We generate a set of data points for each experiment with about half for treated, half for the control group. We are interested in comparing the accuracy of the four different causal inference techniques. We measure the MSE between estimated $\hat{\tau}(x)$ and the ground truth of $\tau(x) = c + \sum_{i=1}^{k} \beta_i x_i$ to evaluate the accuracy of the estimates. Note that this is not the artificial best case configuration, so the data set does not contain exactly matching pairs of treated and untreated units. All forests are computed with 1000 trees.

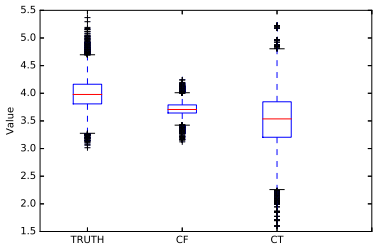


$\mathcal{N}(0,1)$	CF	CT	TOT	RFTOT
min	1.70E0	2.56E0	1.28E2	4.15E1
max	1.76E0	2.76E0	1.31E2	4.27E1
mean	1.72E0	2.68E0	1.29E2	4.20E1
var	6.0E-4	6.0E-3	2.43E0	2.5E-1
fitted	CF	CT	TOT	RFTOT
min	5.7E-2	2.9E-1	5.13E2	1.71E2
max	6.3E-2	3.3E-1	5.46E2	1.82E2
mean	6.9E-2	3.2E-1	5.22E2	1.74E2
var	5.4E-6	1.9E-4	1.8E2	2.1E1

Figure 3: Estimated treatment effect distributions for a single experiment with sampling from fitted distributions, MSE statistics for repeated runs for covariates sampled from normal and fitted distributions.

The box plots in Figure 3 show the distribution of $\tau(x)$ values for the ground truth and estimates $\hat{\tau}(x)$ for a causal forest CF and a causal tree CT for the first scenario. The box plot for the ground truth (TRUTH) shows that the true causal effect values vary only very lightly on the given scale, which is plausible as only a single random variable creates some variability for it. While CF and CT are on the right scale, none of them have an interquartile range that matches well with the true distribution. CF notably underestimates the variability in $\tau(x)$. The box plots do not include results for TOT and RFTOT as their box plots spread wider by orders of magnitude and distort the visualization. This is easier to see in the numerical values in the table of Figure 3 which reports some basic statistics on MSE values (and not treatment effects) seen across a repeated set of 5 experiments for the scenario with sampling from a normal distribution on top and from the fitted distributions on the bottom. The sharp increase in d made it more difficult to estimate $\tau(x)$ for the chosen value of n. For the normal distribution scenario, we can see that results for CF and CT are in a range that one could expect to achieve reasonably accurate results for some higher settings of n if one takes trends into account that are seen on the right side of Fig. 2. TOT and RFTOT perform one to two orders of magnitude worse than CF and CT, which is consistent with what we saw in the previous experiment (graph in the middle of Fig. 2). Using fitted distributions instead of normal distributions makes these issues even more pronounced. MSE values for CF and CT improve for fitted distributions compared to normal distributions but get worse for TOT and RFTOT. We attribute the improvement for CF and CT in part to the feature scaling that limits the range of covariates to $[0,1]$ such that $\tau(x) \geq 1$. As TOT and RFTOT are clearly underperforming, we focus our analysis on CF and CT for the following.

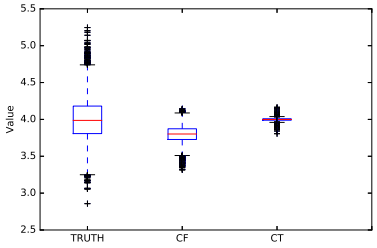
Looking at a single covariate to determine the causal effect is a particular corner case. Figure 4 shows results for the same model configuration and for a single experiment for increasing values of $k = 1$, 2, 4, and 8. The values in the table show that MSE values tend to remain in the same order of magnitude. For both distribution scenarios, values increase with increasing values of k. This means when there are



	$\mathcal{N}(0,1)$		fitted	
k	CF	CT	CF	CT
1	1.69	2.69	0.05	0.31
2	2.66	3.73	0.07	0.35
4	4.63	5.82	0.09	0.40
8	8.42	9.98	0.11	0.49

Figure 4: Estimated treatment effects for CT and CF for $k = 8$ and MSE values for varying values of k.

more confounders in the model, it is harder to estimate the causal effect. The box plots in Figure 4 show causal effect values for the case of sampling from fitted distributions and $k = 8$ and essentially match with what we have seen in Fig. 3 before, but for a ground truth that has more variability that is not adequately matched by CF or CT results. However, CF consistently achieves better MSE values than CT. For the scenario with fitted distributions, the rightmost columns in the table in Fig. 4 show small error values as before, which seem less affected by k than corresponding results for the Normal distribution case.



	$\mathcal{N}(0,1)$		fitted	
k	CF	CT	CF	CT
1	1.63	0.99	0.03	0.007
2	2.57	2.00	0.04	0.02
4	4.42	3.96	0.05	0.04
8	8.16	7.91	0.07	0.08

Figure 5: Estimated treatment effects for CT and CF for $k = 8$ and MSE values for varying values of k for the artificial best case scenario with matching pairs of data points for treated and untreated cases.

The estimate of $\hat{\tau}(x)$ is based on the comparison of specific subsets of treated and untreated units. The accuracy of the outcome is influenced by the quality of the data set which can provide treated and untreated units that are more or less comparable or similar to each other. In order to see what CT and CF can achieve if applied to high quality data, we provide an artificial best case. We repeat the experiments with the same parameter settings and ranges but produce exact pairs of treated and untreated units. Of course, this is an impossible best case where the causal treatment effect calculation is trivial, but the question is if CT and CF approaches benefit from this. The question behind this is to what extent the binning that the tree calculation performs in its leaf nodes introduces errors into this causal effect calculation. Figure 5 shows corresponding results that we can directly compare with Figure 4. For CF, the box plot with $\hat{\tau}(x)$ estimates and the MSE values for the $\mathcal{N}(0,1)$ case do not change significantly. When sampling from the fitted distributions, we see a moderate reduction in the MSE values for CF but a significant reduction for CT that makes CT even more accurate than CF (with the exception of $k = 8$). The box plot shows that CT aligns well with the center of the true distribution but underestimates the interquartile range and variance. The CT approach benefits from this best case scenario. The MSE values for CT are slightly better than the ones for CF for both sampling distributions (with the exception of $k = 8$ for the fitted case). We see two main reasons for the CF to essentially retain its MSE values: the generation of a causal forest considers random subsets of values, which can break up the perfect pairs of data points we provide. Secondly, having a minimum set of treated and untreated cases in each leaf node will imply that even if perfectly matching pairs are present in a leaf node, the treatment effect is calculated with average values that may also include data points without a matching unit. So, if we only subsample the treated units, this achieves better results.

Table 1: MSE results for different configurations of the splitting rule.

	Artificial Best Case				Unpaired Random Samples			
$\mathcal{N}(0,1)$	Min=2		Min=10		Min=2		Min=10	
k	CT	CF	CT	CF	CT	CF	CT	CF
1	0.00	1.24	0.99	1.64	10.71	1.83	2.71	1.65
2	0.00	1.83	1.97	2.56	11.86	2.62	3.60	2.65
4	0.00	2.87	3.90	4.39	14.36	4.17	5.77	4.55
8	0.00	4.58	8.06	8.31	19.46	7.26	10.33	8.53
fitted	Min=2		Min=10		Min=2		Min=10	
k	CT	CF	CT	CF	CT	CF	CT	CF
1	0.00	0.11	0.01	0.03	0.41	0.22	0.33	0.06
2	0.00	0.12	0.01	0.03	0.43	0.23	0.34	0.06
4	0.00	0.13	0.03	0.05	0.50	0.28	0.42	0.08
8	0.00	0.18	0.09	0.08	0.58	0.34	0.50	0.12

Comparing results in Figs. 4 and 5 shows that a high quality data set with exactly matching pairs of treated and untreated cases does not lead to perfect results being calculated by the CF. We find that the required minimum leaf size is a parameter that also influences the quality of the estimated treatment effect. For the artificial best case, a minimum leaf size of two is a promising candidate. So, we exercise these experiments again but this time adjust the parameter for the splitting rule to allow for a single treated and a single untreated case in a leaf node. The latter case is denoted as $Min = 2$ in Table 1, while columns under $Min = 10$ denote that the splitting rule requires a total of 10 elements in a leaf node with at least 1 of each kind (treated or untreated). The results for the $\mathcal{N}(0,1)$ case show that CT can get exact results for a leaf size of 2 in the artificial best case. However, there are opposite trends: if one moves from the best case scenario to unpaired random samples, i.e. one reduces the quality of the data set, then increasing the minimum leaf size from two to ten helps the CT approach while for the best case scenario on its own, an increase in the leaf size increases MSE values. For the $\mathcal{N}(0,1)$ case, the CF approach is rather insensitive to changes of the minimum leaf size which suggests smaller value settings to include the corner case of the best case scenario.

If we sample data points from distributions fitted to the world bank data, we see that the CT approach for the artificial best case and the minimum leaf size of two is accurate. An increase of the leaf size for the best case scenario introduces errors for CT. However, for the regular case of unpaired random samples, a minimum leaf size of ten slightly reduces MSE for CT. For the CF approach, one can see that a minimum leaf size of ten in general is better than two regardless of the presence of perfectly matching data points or not.

4.2 Convergence Rate For Increasing Data Size

Of course, one can not expect to find perfectly matching pairs in the data set as for our artificial best case scenario. It is more realistic to assume that with an increasing size of the data set, we will see more units that are similar in their covariate values and thus expect that the quality of the approximation of $\tau(x)$ improves with n, the size of the data set. Following this, we also test the convergence of the causal inference methods as the size of data increases, as shown in Figure 6.

We consider a model configuration with $k = 1$, $d = 37$, $c' = 1$ for increasing values of n. The minimum leaf size is 10. We sample covariates from fitted distributions. This is not the artificial best case scenario. This experiment essentially tries to achieve the observed decrease in MSE values for increasing n that we have seen for the lower dimensional case in Fig. 2 for a model configuration with more covariates and sampling from different distributions. Figure 6 shows the MSE of CT and CF for a series of experiments

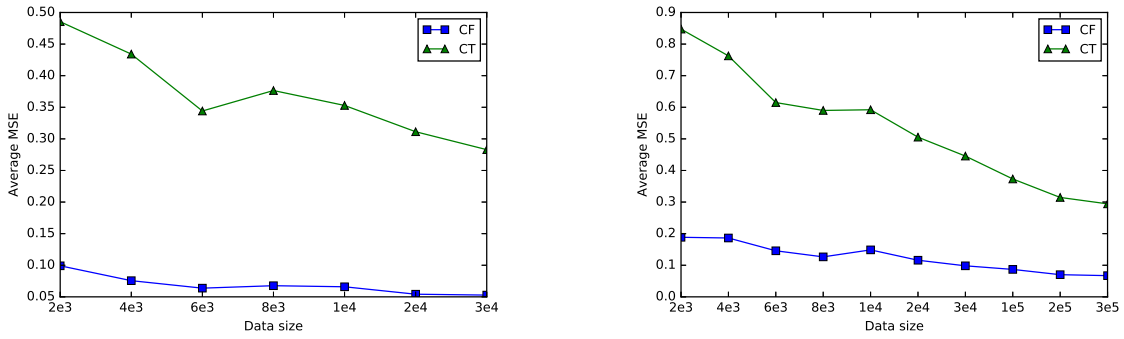


Figure 6: MSE changes with data size for CT and CF, left figure k=1 and right figure k=8.

with increasing data size that starts with $n = 2000$ and increases n by more than one order of magnitude. The observed results do show a pronounced underlying trend and confirm observations for the lower dimensional case.

5 DISCUSSION AND CONCLUSIONS

In this paper, we used a stochastic input model to evaluate the ability of tree-based causal inference techniques to accurately estimate a heterogeneous treatment effect. We contrasted four different approaches all based on variations of regression trees and random forests of trees: Transformed Outcome Trees (TOTs), Causal Trees (CTs), Random Forest TOTs (RFTOTs), and Causal Forests (CF). We found that the method selected can have significant influence on the causal effect (or lack thereof) estimated, and provide evidence suggesting CF is more accurate than alternatives in our study context. The conducted simulations helped us to overcome the challenge that the ground truth in causal inference is not known and that for a specific unit one cannot observe both outcomes for treated and control at the same time.

As we are interested in applying these techniques to calculate the impact of World Bank aid projects on the environment as measured by an index for vegetation cover (NDVI), we configured our stochastic input model to produce covariate data of similar kind. We assumed independence of covariates such that we could fit a distribution for each individual covariate and sample from that distribution. We calculated the correlation between covariates in the given real data set and found very modest correlations among most covariates, such that this assumption seems reasonably satisfied. A much stronger assumption is the additive model that we used to compute an outcome Y that is in turn used to calculate treatment effects. While the covariates in the data set are reasonable - precipitation, temperature, population density relate to vegetation cover - we do not have an established functional relationship. This necessarily limits the scope of this investigation to be relevant primarily for aspects of systems which can be modeled using linear approximations; future work into simulating datasets with non-linearities is ongoing. Errors increase with the number of covariates d and with the subset of covariates k that are responsible for the treatment effect. Also the selection of distributions used to sample covariate values and their scaling influenced the accuracy in our experiments. An increase in the data size n helped to reduce the MSE, which is as expected. As our artificial best case of perfectly matching pairs demonstrated, one can also look for ways to obtain more similar or better pairs to match than just obtain more data. Our simulation results clearly show differences in the achieved accuracy if one replaces sampling from a normal distribution that allows treatment effects to be positive or negative with sampling from fitted distributions which gives more realistic covariate values for our intended application data set. For the latter case, sample covariate values are scaled for the calculation of outcome Y which implied that treatment effects can be more or less positive but not negative. We consider this one of the main reasons for the observed improvement of accuracy.

There are a number of issues: at a very technical level for the splitting rule of trees, what is the best way to select proper limitations on the makeup of terminal nodes - i.e., if splits that result in nodes without both

control and treatment cases should be prevented, omitted, or otherwise constrained. Even after propensity score adjustments, terminal nodes with no adequate comparison cases lack a well-defined interpretation.

REFERENCES

Athey, S., and G. Imbens. 2015. "Recursive Partitioning for Heterogeneous Causal Effects". *arXiv:1504.01132*.

Biau, G. 2012, April. "Analysis of a Random Forests Model". *JMLR* 13 (1): 1063–1095.

Breiman, L. 2001, October. "Random Forests". *Mach. Learn.* 45 (1): 5–32.

Denil, M., D. Matheson, and N. de Freitas. 2014. "Narrowing the Gap: Random Forests In Theory and In Practice". In *ICML*.

Hirano, K., G. Imbens, and G. Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score". *Econometrica* 71 (4): 1161–1189.

Ho, D., K. Imai, G. King, and E. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference". *Political Analysis* 15:199–236.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Meinshausen, N. 2006, December. "Quantile Regression Forests". *JMLR* 7:983–999.

Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika* 70:41–55.

Su, X., C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. 2009. "Subgroup Analysis via Recursive Partitioning.". *Journal of Machine Learning Research* 10:141–158.

Tibshirani, R. 1994. "Regression Shrinkage and Selection Via the Lasso". *Journal of the Royal Statistical Society, Series B* 58:267–288.

Vapnik, V. N. 1998. *Statistical Learning Theory*. Wiley-Interscience.

Wager, S., and S. Athey. 2015. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *arXiv:1510.04342*.

Wager, S., T. Hastie, and B. Efron. 2014, January. "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife". *JMLR* 15 (1): 1625–1651.

AUTHOR BIOGRAPHIES

JIANING ZHAO is a PhD candidate of Computer Science at the College of William and Mary. His research interests lie in simulation modeling, data mining, causal inference especially in social good. His email address is jzhao@cs.wm.edu.

DANIEL RUNFOLA is a Research Assistant Professor at the College of William and Mary (previously CU Boulder and the National Center for Atmospheric Research). His research interests include computational methods for large-scale analyses of geospatial data, the evaluation of international aid effectiveness, and the application of HPC simulation techniques to policy relevant issues. He contributed to the selection and construction of the dataset, the identification of problem sets, and the application of Causal Tree analyses. His email address is drunfola@aiddata.org.

PETER KEMPER is an Associate Professor in the Department of Computer Science at the College of William and Mary. His research interests include modeling techniques and tools for performance, performability and dependability analysis of systems. He contributed to analysis techniques for the numerical analysis of Markov chains, model checking stochastic models, techniques for simulation optimization. His email address is kemper@cs.wm.edu.