Karlsruher Institut für Technologie

Department of Economics and Management

Institute of Economics (ECON)

Chair of Econometrics

Prof. Dr. Melanie Schienle

Bachelor's Thesis

# Machine Learning for Causal Inference: Estimating Heterogeneous Treatment Effects with Causal Forests

by

Frederike Lübeck

2055292

Industrial Engineering and Management (Bachelor)

Date of Submission

May 28, 2020

## Abstract

There is a growing interest in evaluating the causal effect of a treatment beyond its average effect. Understanding how a treatment varies among sub-populations is a key question across research fields ranging from medicine to policy. This understanding is important to enable personalized treatments or well targeted interventions. Due to the large amount of data that is constantly produced, there is a demand for methods that automatically detect heterogeneity in treatment effects. Since machine learning has shown to perform well on similar tasks, it is promising to apply machine learning based methods to causal inference. However, most methods are not directly applicable and need to be tailored to the task of estimating treatment effects.

This work analyzes the recently introduced Causal Forests in the framework of Generalized Random Forests of Athey et al. (2019), which is based on the popular Random Forest algorithm of Breiman (2001). The theoretical part of this thesis provides a comprehensive introduction into tree-based methods and their application to estimating treatment effects. On simulated data, I demonstrate that Causal Forests are able to automatically detect strong treatment effect heterogeneity. Further, I show that the challenge of confounding is effectively addressed by the technique of local centering. Investigating the implications of the underlying concept of *honesty*, I find that Causal Forests successfully resist overfitting to noise. Lastly, I apply the Causal Forest algorithm to data from a field experiment in political science from Broockman and Kalla (2016), who show that brief but high-quality canvassing can markedly reduce prejudice against transgender people. Overall, the results using Causal Forests confirm the findings of the authors. Only small variation in treatment effects, but no statistically significant heterogeneity is found. Simultaneously, an approach to make best use of Causal Forests is demonstrated in order to enable researchers outside the field of statistics to take advantage of this machine learning method in the setting of estimating treatment effects.

# Contents

## List of Abbreviations

| | |
|---|---|
| ATE | Average Treatment Effect |
| CART | Classification And Regression Tree |
| CATE | Conditional Average Treatment Effect |
| CF | Causal Forest |
| GMM | Generalized Method of Moments |
| GRF | Generalized Random Forests |
| kNN | k-Nearest-Neighbors |
| MSE | Mean Squared Error |
| PAP | Pre Analysis Plan |

# List of Figures

# List of Tables

# 1 Introduction

Measuring the effectiveness of a treatment or program is of considerable interest in a variety of research disciplines and applications. In medicine, for example, the effect of a drug on a patient's health must be well-known before being prescribed. Yet, a drug that cures some people, might harm others, depending on a patient's medial history. Thus, beyond determining whether a treatment works on average, it is important to understand treatment effect heterogeneity and to identify sub-populations for which the effects differ. There is a large number of current interesting applications, ranging from personalizing medicine to optimally allocating a scarce resource and targeting in marketing or policy campaigns (Ascarza, 2018; Brand and Xie, 2010; Obermeyer and Emanuel, 2016; Kravitz et al., 2004). In these areas, new data is constantly collected and this opens up possibilities for making decisions more data-driven.

The ability to transform data into knowledge depends on using appropriate methods. In recent years, attention has shifted to new statistical tools from the field of supervised machine learning. A large number of machine learning algorithms have become popular for successfully learning how to predict an outcome based on data. These methods *learn* a regression function on their own by looking at training data and searching for combinations that reliably predict outcomes. At each step, a model's predictive power is evaluated by comparing the predictions to the observed outcomes. This way, the best estimator is found. This data-driven approach has proven powerful, especially in high-dimensional data sets in which the number of variables is higher than the number of observations.

However, estimating the causal effect of a treatment is different from predicting an outcome. Here, the problem is that the ground truth of a causal effect is never observed. Since each individual either received the treatment or didn't, we are not able to observe the counterfactual outcome that would have resulted from a different treatment. Thus, it is not obvious how to determine whether a causal effect has been accurately predicted, yet this is necessary for selecting the best estimator.

Further, empirical researchers are typically not only concerned with predictions, but seek to draw statistical inference about the magnitude of treatment effects. When applying machine learning methods *off-the-shelf*, without understanding their theoretical properties and without guaranteeing that the required conditions are met, conclusions are questionable. Researchers need to make sure to detect causality, not simply correlation. In the causal setting, there is a need to interpret results. Understanding the reasons for treatment effect

heterogeneity is particularly crucial if decisions are made based on data, such as allocating resources. There is a difference in making a prediction and making a decision, and this needs to be taken into account when developing methods.

To benefit from the advantages of machine learning when estimating treatment effects, these algorithms need to be modified. In recent years the number of new methods that take up the challenge of making machine learning *causal* has grown. Nonetheless, these numerous papers do not yet provide a comprehensive answer on how to use machine learning for estimating causal effects. This lack of comprehension can seem daunting for researchers outside the field of statistics, even though machine learning methods can substantially improve upon traditional procedures.

In this thesis, I analyze a recently introduced method of Athey et al. (2019), namely *Generalized Random Forest*. Building up on a large literature on tree-based methods, they propose a causal variant of Random Forests. The focus of this work is to discover the benefits this algorithm brings to the task of estimating heterogeneous treatment effects. By investigating how Causal Forests perform in different scenarios, I provide intuition on how — and when — to use them. I show how to interpret results and identify subgroups with distinct treatment effects. Analyzing an empirical data set of the field of political science, I demonstrate how to make best use of Causal Forests.

The rest of this thesis is organized as follows. In Section 2, I set up the framework for estimating treatment effects and point out the challenges that arise. Section 3 yields a comprehensive introduction into tree-based methods and Generalized Random Forests in particular. After explaining their statistical properties and algorithmic procedure, I discuss advantages over traditional methods. In Section 4, I conduct an extensive simulation study showing the performance of Causal Forests in a variety of scenarios. An empirical application in which I make use of Causal Forests is presented in Section 5. In Section 6, I discuss the results and finally, in Section 7, I draw a conclusion on the benefits of Causal Forests.

# 2 Treatment Effect Estimation

This Section aims to provide the theoretical background necessary in order to define the requirements that the estimator used in Section 3 should meet. In order to achieve this aim, I start with introducing the basic terminology with the help of the potential outcomes framework and move on to explaining the assumptions needed. Finally, I present the estimand of interest and define the requirements.

## 2.1 Potential Outcomes Framework

In many applications, economists and other professionals are interested in measuring the effectiveness of a program, a policy or a treatment. Suppose we observe $N$ units, indexed by $i = 1, ..., n$, viewed as drawn randomly from a large population. Each unit is either exposed to an active treatment or to a control treatment, e.g. a placebo. This binary treatment indicator is denoted by

$$W_i = \begin{cases} 1, & \text{if unit } i \text{ receives the active treatment} \\ 0, & \text{if unit } i \text{ receives the control treatment.} \end{cases}$$

In addition, we observe a potentially high-dimensional vector of characteristics for each unit, referred to as covariates and denoted by $X_i$. These covariates can be seen as *pre-treatment* characteristics and are not affected by the treatment. Finally, for each unit we observe an outcome $Y_i$. Hence, our data consists of the triple $(X_i, W_i, Y_i)$ for $i = 1, ..., n$.

Following the Potential Outcomes Framework developed by Rubin (1974), we postulate the existence of two potential outcomes for each unit. $Y_i^{(1)}$ denotes the outcome if unit $i$ receives the active treatment, and the counterfactual outcome $Y_i^{(0)}$ if unit $i$ would have received the control treatment, respectively. Then, the treatment effect $\tau$ for each unit is the difference between the two potential outcomes:

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

Our goal is to estimate this function $\tau(x)$. Although each unit is associated with two potential outcomes, only one of them can be observed, depending on which treatment the unit actually receives. It is impossible to observe the causal effect on a single unit and therefore, we cannot directly train a model on the differences of the form as in (1). Holland (1986) refers to this as the fundamental problem of causal inference. To still identify the causal effect of a treatment, further restrictions are needed.

## 2.2 Assumptions

A simple starting point to evaluate a treatment program would be to compare the average outcomes of the two treatment groups and to interpret the difference as the average treatment effect. While this might work in a randomized experiment, this might lead to biased estimated in observational studies. In a randomized experiment, the treatment is randomly assigned to the units. Thus, the distribution of the covariates $X$ is assumed to be the same in both treatment groups and comparing both groups will then lead to an unbiased estimate of the average treatment effect $\tau$. However, in observational studies, the treatment is not necessarily randomly assigned. Often some covariates, referred to as confounders, affect both the treatment assignment $W_i$ and the outcome $Y_i$. Simply comparing the outcomes of both groups can lead to biased estimates.

To illustrate this, imagine we want to measure the effect of a drug on a patient's health. For some unknown reason, younger patients are more likely to take the drug. Therefore, the average age in the treated group is lower than in the control group. Simply comparing the health of the two groups to identify the effect of the drug would lead to a biased estimate, since it might be that younger people are generally healthier than older people. We need to account for this systematic differences between the groups.

In order to adjust for differences in these exogenous background characteristics, we need to be able to observe them. Thus, we make the following key assumptions.

**Assumption 1** (Unconfoundedness)**.**

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp\!\!\!\perp W_i \mid X_i$$

The Unconfoundedness-Assumption states that the treatment assignment is independent of the potential outcomes conditional on X. Put differently, this means that we observe all covariates that confound the treatment assignment. The potential outcomes only depend on these covariates.

In addition, another assumption regarding the joint distribution of treatments and covariates is needed.

**Assumption 2** (Overlap)**.**

$$0 < \mathbb{P}\left[W_i = 1 \mid X_i = x\right] < 1$$

The treatment assignment is not deterministic for each set of values for $X$, that means that the conditional probability of receiving the treatment is bounded between zero and one.

Thus, everyone in the population could possibly receive the treatment. This conditional probability of receiving the treatment given the covariates is known as the propensity score and is denoted by $e(x)$. In a randomized experiment with constant treatment assignment, the propensity score is a known value that is equal for all values of $x$. In nonrandomized trials, the propensity score function is almost always unknown (Rosenbaum and Rubin, 1983).

Rosenbaum and Rubin (1983) refer to the combination of assumptions 1 and 2 as *strong ignorability*. Unconfoundedness implies that we can treat nearby observations as if they came from a randomized experiment, thus, that we can estimate the treatment effect by comparing the outcomes of nearby observations. In addition, the Overlap-Assumption guarantees that, for large enough $N$, there will be enough treatment and control units near any test point x for local methods to work. Given these assumptions, we can identify the average treatment effect.

Furthermore, we assume that the Stable Unit Treatment Value Assumption (Rubin, 1980) holds. This ensures that the treatment assignment of one unit does not affect the outcome and treatment assignment of another unit.

### 2.3 Estimands

The most commonly studied estimand in the econometric literature is the population average treatment effect (ATE), which is defined as

$$\tau = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)}\right]. \tag{2}$$

However, treatment effects potentially vary across covariates. Therefore, to understand heterogeneity in treatment effects we mainly focus on accurately estimating the conditional average treatment effect (CATE), defined as

$$\tau(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \mid X_i = x\right]. \tag{3}$$

The expectation of CATE estimates over all values of $x$ is similar to the ATE.

Given the before-mentioned assumptions, the key insight is that the following equations hold:

$$\mu_w(x) = \mathbb{E}\left[Y_i(w) \mid X_i = x\right] = \mathbb{E}\left[Y_i \mid W_i = w, X_i = x\right] \tag{4}$$

where $\mu(x)$ denotes the conditional expectation of $Y_i$. Thus, the treatment effect is identified

from observable data by

$$\begin{aligned}\tau(x) &= \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = x\right] \\ &= \mathbb{E}\left[Y_i(1) \mid X_i = x\right] - \mathbb{E}\left[Y_i(0) \mid X_i = x\right] \\ &= \mathbb{E}\left[Y_i \mid W_i = w, X_i = x\right] - \mathbb{E}\left[Y_i(0) \mid W_i = w, X_i = x\right].\end{aligned} \tag{5}$$

The fundamental task is to estimate these conditional expectations.

### 2.4 Challenges and Requirements

Detecting heterogeneity in treatment effects has important implications in applications ranging from medical research to social and political science (Brand and Xie, 2010; Kravitz et al., 2004; Obermeyer and Emanuel, 2016). In medicine, recognizing heterogeneity in the effect of a drug is essential in order to solely prescribe it to patients that benefit from it. Regarding policy decisions, finding those individuals that benefit most from an intervention helps allocating a costly or sparsely available treatment. Traditionally, heterogeneity estimation often involves specifying the subgroups between which a difference is suspected in advance, e.g. in a pre-analysis plan. The reason for this is to avoid concerns about multiple testing (Athey and Imbens, 2015), whereby the researcher conducts a large number of hypothesis tests until some heterogeneity is found, even if it might be purely spurious. In a second step, the sample is divided into the subgroups and a separate model is trained on each one. However, with such procedural restrictions, researchers might miss unexpected heterogeneity. Thus, we seek a method that lets the data discover relevant subgroups while still preserving the validity of the results. In addition, we want to interpret the differences by analyzing the baseline covariates of the subgroups that affect the treatment effect.

Secondly, estimating treatment effects often occurs in settings in which rigorous statistical inference is required. Therefore, a consistent estimate of $\tau(x)$ and its variance is required in order to accurately construct valid confidence intervals.

Thirdly, our method needs to adequately adjust for confounders in observational studies. Hence, our interest is in a method that can handle both, treatment effect heterogeneity and confounding effects.

Finally, in the spirit of Big Data, we seek a method that performs well even when the number of attributes is much larger than the sample size. This is particularly interesting because many traditional methods such as nearest-neighbor matching or propensity score weighting are prone to a strong curse of dimensionality.

# 3 Generalized Random Forests

Generalized Random Forests, a method proposed by Athey, Tibshirani, and Wager (2019), seems highly promising to tackle the before-mentioned challenges. This section yields a comprehensive introduction into their method. To outline the progression of Generalized Random Forests, I start with explaining the traditional tree and forest algorithms. I then show difficulties that arise when using Random Forests for estimating treatment effects. Finally, I present how Athey et al. (2019) modify Random Forests into Causal Forests.

## 3.1 Breiman's Algorithm: From Trees to Forests

The general framework in which tree-based methods are used is non-parametric regression estimation. Accordingly, we observe an input vector $X \in \mathcal{X} \subset \mathbb{R}^p$ and the goal is to predict the response $Y \in \mathbb{R}$ by estimating the regression function

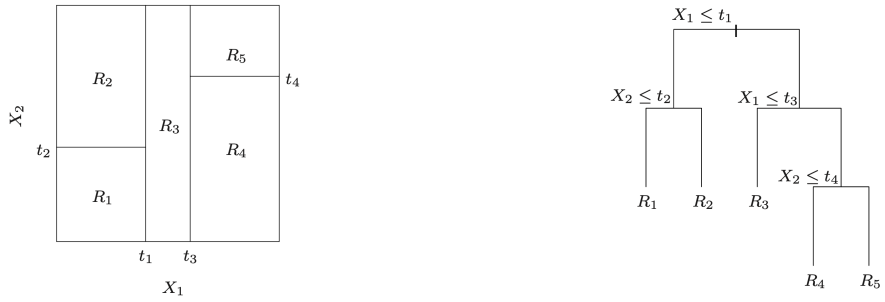
$$\mu(x) = \mathbb{E}\left[Y \mid X = x\right] \tag{6}$$

using a training sample $\mathcal{S}_{tr} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$. We will modify this framework in Section 3.2 to estimate treatment effects.

A tree recursively partitions the feature space $\mathcal{X}$ into a set of rectangles and then fits a simple model (e.g. a constant) in each region. See Figure 1 for an illustration of such a partition, presented in form of rectangles as well as a binary tree. Splits are placed in order to minimize the impurity in the resulting regions. In the traditional Classification and Regression Trees (CART) algorithm proposed by Breiman et al. (1984), the impurity is defined as the mean squared error (MSE), that is $\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\mu}(x_i))^2$. Thus, in each region, the MSE-minimizing prediction corresponds to the mean of the responses of observations that fall into that region. We seek to find partitions, denoted by $\Pi$, that minimize the impurity criterion or – equivalently – maximize the in-sample goodness-of-fit.

However, since finding the best binary partition in terms of minimum sum of squares is computationally infeasible, the algorithm proceeds greedily. At each node, a split point is chosen in order to maximize the model fit as fast as possible. This means, the algorithm evaluates the goodness-of-fit criterion over all possible axis-aligned splits and returns the best one. This process is then repeated recursively in the two corresponding child nodes. Tree construction is stopped when a pre-defined condition is met, such as when a leaf contains less than a certain number of observations.

A tree's complexity is determined by its size. While a large tree might overfit to the




**(a)** rectangles

**(b)** binary tree

**Figure 1:** A simple example of recursive partitioning a two-dimensional feature space by adding axis-aligned splits. Figure from (Hastie, Tibshirani, and Friedman, 2008)

data, a small tree might not capture the important structure. Typically, in a standard CART routine, trees are grown until a pre-defined minimum node size is reached. Then, the large tree is pruned by collapsing some of the internal nodes if the improvement in model fit due to this split does not exceed a certain threshold. This is accomplished using Cross-Validation, whereby the predicting performance is evaluated by leaving out the test points when predicting their outcome. Thus, the aim of pruning is to find a good tradeoff between a tree's complexity and it's goodness of fit to the data.

As a result, when the tree construction is done, the feature space is partitioned into a set of terminal leaves $L$, each of which contains some of the training observations. Given a test point $x$, we predict the response by identifying the leaf $L(x)$ containing $x$ and by estimating the response as the mean in that leaf, that is

$$\hat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{\{i:X_i \in L(x)\}} Y_i. \tag{7}$$

A major problem of individual regression trees is their high variance. Small changes in the training data can result in different splits and therefore in very different trees. This is due to the fact that an error made in a split in the top of the tree is propagated down to all splits below (Hastie, Tibshirani, and Friedman, 2008). A solution to this problem, proposed by Breiman (2001), is to generate many trees and to combine them into a forest. As shown by Bühlmann and Yu (2002), this aggregation scheme reduces variance and therefore stabilizes predictions. Combining trees is based on a technique called Bagging or *Bootstrap Aggregation*. Each tree is grown on a different subsample randomly drawn from the original data set and considers only these observations in the tree-building process. After $B$ such trees are grown, the Random Forest predictor is then obtained by averaging the predictions of all trees, that

is

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_b(x). \tag{8}$$

Random Forests can be seen as a form of nearest-neighbor estimator in that they make predictions using a weighted average of nearby observations in the $x$-space. However, Random Forests are an adaptive version of nearest-neighbor matching for having a data-driven way to determine which observations are "nearby". By splitting on variables that have great influence on the outcome, each tree defines which variables should be used to describe the neighborhood, and which variables should be left out of that description, respectively. Being able to ignore variables that do not affect the outcome, Random Forests are generally recognized for their ability to handle high-dimensional feature spaces. This key advantage is strongly contributing to their popularity.

Since their introduction in 2001, Random Forests have proven to be a simple yet powerful method for non-parametric regression and classification. Because forests can be applied to a wide range of statistical prediction problems, they receive widespread attention in various research areas ranging from econometrics to bioinformatics.

### 3.2 Causal Forests: Forests for estimating treatment effects

Although predicting conditional expectations and predicting treatment effects are closely related problems, Random Forests cannot be applied off-the-shelf to the problem of treatment effect estimation. The setting with treatment effects comes with some specific problems.

First, if instead of $\mu(x)$ we want to predict $\tau(x)$, we face the challenge that the ground truth of a treatment effect is never observed. Therefore, it is infeasible to directly evaluate splits in terms of the difference between the true treatment effect and the estimate, $(\hat{\tau}(x) - \tau(x))$ (Athey and Imbens, 2016). Nonetheless, this is an essential part of the original Random Forest algorithm to evaluate possible split points. Hence, a new criterion is required that consistently estimates the goodness-of-fit.

Second, the aim is to draw valid statistical inference about the magnitude of treatment effects. Despite their widespread practical application, only little has been known about the theoretical and mathematical properties of Random Forests (Wager and Athey, 2018; Biau and Scornet, 2016). Ideally, an estimator should be consistent for the true treatment effect with a well understood asymptotic sampling distribution (Wager and Athey, 2018). Thus, in order to use Random Forests for causal inference, their asymptotic properties need to be known.

Third, as mentioned in Section 2, we seek to find heterogeneities in treatment effects. Thus, the success of Random Forests depends on whether the partitioning adequately captures the heterogeneity in $\tau(x)$ (Athey et al., 2019). If we would, for example, predict the outcome $Y$ with a forest and then, in a second step, compare the outcomes between treated and control units, we might not find all heterogeneities in treatment effects. The reason for this is that a covariate that affects the expected outcome $\mathbb{E}[Y|X = x]$ does not necessarily affect the treatment effect $\mathbb{E}[\tau|X = x]$, and vice versa. Therefore, trees might not split on the covariates that cause treatment effect heterogeneity. The goal is to tailor the splitting to finding heterogeneity in the parameter that we are interested in.

Despite these challenges, the advantages of recursive partitioning and the forest aggregation scheme are still promising for estimating heterogeneous treatment effects. Therefore numerous researchers have focused on developing new methods that can preserve these advantages but also solve the aforementioned challenges. For example, Su et al. (2009) proposes a partitioning algorithm for subgroup analysis and Hill (2011) suggest using Bayesian Additive Regression Trees for causal inference. Zeileis et al. (2008) proposed to fitting a parametric model in each node and to performing a test for parameter instability to determine whether a split is necessary. In addition, Denil et al. (2014), Biau (2012) and Wager et al. (2014) analyzed the asymptotic properties of forests.

Particularly interesting for this thesis is the sequence of papers of Athey and Imbens (2016), Wager and Athey (2018) and Athey, Tibshirani, and Wager (2019). For being powerful yet theoretically well understood, these algorithms seem exceptionally promising for treatment effect estimation.

Athey and Imbens (2016) proposed *Causal Trees*, which use a new splitting rule tailored to estimating treatment effects. In a subsequent paper, Wager and Athey (2018) extended these trees into a *Causal Forest*. Both also developed a large sample theory about the asymptotic properties of their algorithm. However, their algorithm has a substantial drawback, since each of the two proposed procedures addresses only one of the challenges, treatment effect heterogeneity and confounding, but not both at the same time.

Finally, Athey et al. (2019) extended its predecessors into a generalized framework. They propose *Generalized Random Forests (GRF)*, which is a unified, flexible framework for estimating heterogeneity in any quantity of interest identified via local moment equations. In particular, they focus on the two settings of quantile regression and treatment effect estimation. In case of estimating treatment effects, they call their algorithm *Causal Forest* as

it replaces its predecessors. According to the authors, GRF is the state-of-the art in this sequence and therefore I mainly focus on Causal Forests in the sense of the GRF paper.

### 3.2.1 Generalized Random Forests

Generalized Random Forests are designed to be an extension of Breiman's algorithm. As original Random Forests usually serve to estimate conditional means, GRF enables to fit any quantity of interest, including treatment effects. Suppose we observe $n$ independent and identically distributed samples, indexed $i = 1, ..., n$. For each sample, we observe a quantity $O_i$ along with a set of covariates $X_i$. $O_i$ encodes problem-specific information relevant for estimating the parameter that we are interested in, $\theta(\cdot)$. In the case of treatment effect estimation, $O_i = \{Y_i, W_i\}$ contains the outcome $Y_i$ and the treatment assignment $W_i$. Given this data, $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$, the target functional $\theta(x)$ is identified via local moment equations of the form

$$\mathbb{E}\left[\psi_{\theta(x),\nu(x)}(O_i) \mid X_i = x\right] = 0 \text{ for all } x \in \mathcal{X} \tag{9}$$

where $\psi(\cdot)$ is some scoring function and $\nu(\cdot)$ is an optional nuisance parameter. Put in words, (9) is a function that scores the parameter value and the observed data, such that their expectation is zero at the parameters true value. Due to the very general characterization of $\theta(x)$, such moment conditions can be used in a wide range of statistical settings, including the estimation of treatment effects.

**Local Maximum Likelihood Estimation**. Following the literature on local maximum likelihood estimation and Generalized Method of Moments (GMM), $\theta(x)$ can be estimated by fitting an empirical version of the estimating equation. The idea is to estimate a parameter's value in a local neighborhood of each point $x$. This is accomplished by assigning weights to each of the training observations according to the relevance for estimating the target at a specific value of covariates $x$. These weights are then used to find good estimates of $\theta(x)$, that is

$$\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \text{argmin}_{\theta,\nu}\left\{\left\| \sum_{i=1}^{n} \alpha_i(x)\psi_{\theta,\nu}(O_i)\right\|_2\right\}. \tag{10}$$

Traditionally, these weights $\alpha_i(x)$ are obtained using a kernel weighting function (Hastie et al., 2008). However, such methods are prone to a strong curse of dimensionality. Therefore, Athey et al. (2019) make use of Random Forests as type of an adaptive nearest neighbor

estimator and propose to obtaining these weights from a forest. This way, the advantage of forests to derive an adaptive neighborhood function is utilized.

**Forest-based weighting**. Generalized Random Forests preserve some core elements of Breiman's algorithm. The observations are recursively partitioned into rectangles and individual trees are combined into a forests using a subsampling mechanism. However, the forest is not used to construct the final estimate, but rather to identify the local neighborhood of each point. For each test point $x$, similarity weights $\alpha_i(x)$ are defined as follows: First, $B$ trees are grown, indexed by $b = 1, ..., B$ and for each such tree, $L_b(x)$ denotes the training examples that fall into the same leaf as $x$. Then, the forest weights $\alpha_i(x)$ are defined as the frequency with which the $i$-th training example falls into the same leaf as $x$:

$$\alpha_{bi}(x) = \frac{\mathbf{1}(\{X_i \in L_b(x)\})}{|L_b(x)|}, \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B}\alpha_{bi}(x) \tag{11}$$

For each test point, these weights sum up to 1. The weights $\alpha_i(x)$ then correspond to the relevance of the $i$-th training observation for computing the treatment effect at $x$.

This weighted set of nearby observations is then used to estimate the parameter of interest locally in the neighborhood of each point. At a high level, the estimates $\hat{\theta}(x)$ are produced by solving (10) with forest-based weights. Consequently, our goal is to find those weights $\alpha_i$ that lead to good estimates of $\theta(x)$.

**Finding the appropriate neighborhood**. We seek to find weights that are as sensitive as possible to heterogeneity in the parameter of interest. In the case of estimating treatment effects, we want to define the neighborhood of any point so that close observations have a similar treatment effect. Thus, in order to receive good weights $\alpha_i(x)$, the splitting scheme is tailored to focus on heterogeneity in the target parameter $\theta(\cdot)$. Given a sample of data $\mathcal{S}$, in each parent node $P \subseteq \mathcal{X}$ we calculate the parameter estimates $(\hat{\theta}_P, \hat{\nu}_P)(\mathcal{S})$ by solving (10) without weights, that is

$$\left(\hat{\theta}_P, \hat{\nu}_P\right) \in \text{argmin}_{\theta,\nu}\left\{\left\|\sum_{\{i\in\mathcal{S}:X_i\in P\}}\psi_{\theta,\nu}(O_i)\right\|_2\right\}. \tag{12}$$

These values then correspond to the "best" parameters in the neighborhood defined as the parent node. When dividing $P$ into two child nodes $C_1, C_2 \subseteq \mathcal{X}$, we want to improve the accuracy of our $\theta$-estimates as much as possible. Correspondingly, we want to minimize the error of our parameter estimates in the child nodes that is defined as

$$\text{err}(C_1, C_2) = \sum_{j=1}^{2} \mathbb{P}\big[X \in C_j \mid X \in P\big] \mathbb{E}\big[(\hat{\theta}_{C_j}(\mathcal{S}) - \theta(X))^2 \mid X \in C_j\big] \tag{13}$$

where $\hat{\theta}_{C_j}(\mathcal{S})$ are fit analog to (12).

However, since we do not observe the true value $\theta$, we cannot compute this criterion. Therefore, Athey et al. (2019) proposed a different characterization of this criterion. Splits are chosen to maximize heterogeneity in the parameter estimates of the child nodes, defined as

$$\Delta(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \Big(\hat{\theta}_{C_1}(\mathcal{S}) - \hat{\theta}_{C_2}(\mathcal{S})\Big)^2 \tag{14}$$

where $n_j$ is the number of observations in node $j$. A split that results in two very different parameters in the children is worth the split, while one that results in almost similar parameters in both children does not improve the purity in the nodes. However, the fraction in the beginning ensures that the split is balanced in terms of the number of observations per child node. Then, $err(C_1, C_2)$ can be written as

$$err(C_1, C_2) = K(P) - \mathbb{E}\left[\Delta(C_1, C_2)\right] + o \tag{15}$$

where $K(P)$ is a deterministic term that measures the purity of the parent node and $o$ is a dominant bias term due to sampling variance.

To provide some intuition on the idea behind this criterion, assume an estimator that minimizes the squared error loss for conditional mean estimation. Given a sample $\mathcal{S}$, the split criterion can be written as

$$\begin{aligned}
\sum_{i \in \mathcal{S}} (Y_i - \hat{\mu}(X_i))^2 &= \sum_{i \in \mathcal{S}} (Y_i^2 - 2 \cdot Y_i \cdot \hat{\mu}(X_i) + \hat{\mu}^2(X_i)) \\
&= \sum_{i \in \mathcal{S}} (Y_i^2 - 2 \cdot (Y_i - \hat{\mu}(X_i)) \cdot \hat{\mu}(X_i) - \hat{\mu}^2(X_i)) \\
&= \sum_{i \in \mathcal{S}} Y_i^2 - \sum_{i \in \mathcal{S}} \hat{\mu}^2(X_i)
\end{aligned}$$

The last equality holds because the average of $(Y_i - \hat{\mu}(X_i)) \cdot \hat{\mu}(X_i)$ is equal to zero, since a tree makes its prediction by averaging outcomes in each leaf. Thus, the squared error loss of a split can be written as $\sum_{i \in \mathcal{S}} Y_i^2 - \sum_{i \in \mathcal{S}} \hat{\mu}^2(X_i)$. The first term, $\sum_{i \in \mathcal{S}} Y_i^2$ is a deterministic term that is the same for all possible splits. Consequently, minimizing the squared error loss is equivalent to maximizing the sum of $\hat{\mu}^2(X_i)$. And this, in turn, is equivalent to maximizing the variance of predictions $\hat{\mu}(X_i)$ for $i \in \mathcal{S}$, since $\sum_{i \in \mathcal{S}} \hat{\mu}(X_i) = \sum_{i \in \mathcal{S}} Y_i$. This way, we can

circumvent the problem that the true treatment effect is never observed. The criterion in (14) emulates this by favoring splits that increase the heterogeneity between the estimates.

**Gradient-based split point detection**. Since optimizing $\Delta(C_1, C_2)$ over all possible axis-aligned splits while explicitly solving for $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ in each candidate child may be computationally expensive, Athey et al. (2019) instead optimize an approximate criterion. They compute gradient-based approximations for the parameters in the child nodes and verify, that the difference between exact and the approximate criterion is within tolerance. For more information on the algorithmic details of GRF see Section 2.3 of Athey et al. (2019).

**Treatment Effect Estimation.** In case of estimating treatment effects, we wish to solve $Y_i = W_i \cdot b_i + \epsilon_i$ by estimating $\beta(x) = \mathbb{E}[b_i | X_i = x]$. The parameter $\beta(x)$ is the conditional average treatment effect (CATE). The scoring function in the local moment equations as in (9) that identify $\beta$, is

$$\psi_{\beta(x),c(x)}(Y_i, W_i) = (Y_i - \beta(x) \cdot W_i - c(x)) \begin{pmatrix} 1 \\ W_i^\top \end{pmatrix} \tag{16}$$

where $c(x)$ is an intercept term.

**Addressing confounding with local centering.** In order to be robust to confounding, Athey et al. (2019) propose to applying an orthogonalization technique. This is done by first regressing out the effect of the features $X_i$ on all outcomes $\{Y_i, W_i\}$ separately. Let $y(x)$ and $w(x)$ be the conditional marginal expectation of $Y_i$ and $W_i$ respectively. Then, centered outcomes are defined as

$$\tilde{Y}_i = Y_i - \hat{y}^{(-i)}(X_i) \text{ and } \tilde{W}_i = W_i - \hat{w}^{(-i)}(X_i). \tag{17}$$

$\hat{y}^{(-i)}(X_i)$ and $\hat{w}^{(-i)}(X_i)$ are leave-one-out estimates, whereby the $i$-th observation is left out of training. The forest is then trained using the centered outcomes $\{\tilde{Y}_i, \tilde{W}_i\}$.

To summarize, the trees in GRF are used to define the neighborhood of each test point. This neighborhood is then used for computing predictions. This forest-based weighting scheme provides some significant advantages. By recursively partitioning the data according to it's covariates, GRF adaptively learns better weights that focus on the problem-specific parameter of interest. The neighborhoods obtained can be narrower along the directions where the signal is changing fast and wider along the other directions.

### 3.2.2 Growing Forests with Theoretical Guarantees: The Concept of Honesty

Athey et al. (2019) show that the estimates obtained via GRF are consistent for $\theta(x)$ and asymptotically Gaussian distributed. In addition, they provide an estimator for their asymptotic variance that enables to construct valid confidence intervals. Thus, GRF can be used for accurate statistical inference. To ensure this, their algorithm relies on a technique they call honesty.

A model is honest, if it does not use the same information for selecting the model structure as for estimating, given a model structure. This is accomplished by splitting the training sample into two distinct samples, one for growing the tree $\mathcal{S}^{tr}$ and the other for estimating $\mathcal{S}^{est}$, given the tree structure. Applying the concept of honesty implies that the asymptotic properties of the estimates within the partitions are the same as if the partition had been exogenously given (Athey and Imbens, 2016).

To illustrate the value of honesty, I revisit an example of Athey and Imbens (2016). Suppose each observation can either have the attribute $L$ or $R$, that is $\mathcal{X} = \{L, R\}$. In this case, there are two possible partitions: $\Pi_N = \{L, R\}$ (no split) or $\Pi_S = \{\{L\}, \{R\}\}$ (full split). Given a sample $\mathcal{S}$, the average outcome in the two subsamples are $\bar{Y}_L$ and $\bar{Y}_R$. A simple tree algorithm is one that splits if the difference in outcomes exceeds a certain threshold $c$:

$$\pi(\mathcal{S}) = \begin{cases} \{L, R\}, & \text{if } |\bar{Y}_L - \bar{Y}_R| \leq c \\ \{\{L\}, \{R\}\}, & \text{if } |\bar{Y}_L - \bar{Y}_R| > c \end{cases} \tag{18}$$

Thus, if we condition on finding that $|\bar{Y}_L - \bar{Y}_R| > c$ in a particular sample, we expect that $|\bar{Y}_L - \bar{Y}_R|$ is larger than the population analog. This might lead to biased estimates.

Instead, placing the splits using a sample $\mathcal{S}^{tr}$, and estimating the effects using an independent sample $\mathcal{S}^{est}$ leads to unbiased estimates. If, in $\mathcal{S}^{tr}$, we find that $|\bar{Y}_L - \bar{Y}_R| > c$, we perform the split. If the same heterogeneity between $\bar{Y}_L$ and $\bar{Y}_R$ is found in $\mathcal{S}^{est}$ as well, everything is fine and we can say that this difference really exists. In contrast, if we do not find any heterogeneity between $\bar{Y}_L$ and $\bar{Y}_R$ in the estimation sample $\mathcal{S}^{est}$, it might have been noise in $\mathcal{S}^{tr}$. Nevertheless, our estimates do not model this noise because they use an independent sample for estimation. To conclude, honesty helps us to model true heterogeneity rather than noise.

The idea of performing the tree construction and the estimation on two independent samples goes back to Biau (2012) and Denil, Matheson, and de Freitas (2014). They propose this sample splitting to show that a Random Forests estimate is consistent. Athey and Imbens

(2016) use this technique for single trees. They claim that although there is a loss of precision due to reducing the sample size by splitting it, there is a benefit in terms of eliminating bias that offsets at least part of the cost. In a simulation study, they show the cost in terms of MSE of honest vs. adaptive estimation. However, Wager and Athey (2018) proposed a forest procedure which enables them to achieve consistency without "wasting" half of the data. Although each observation can only be used for split selection or leaf estimation in a single tree, it can participate in both samples in different trees. By re-randomizing the $\mathcal{S}^{tr}/\mathcal{S}^{est}$-data split in each tree, each observation is used for both, tree construction and leaf estimation. In Section 4, I demonstrate that honest GRF can improve upon adaptive GRF in terms of MSE in my simulations. In addition, I illustrate that the risk of modeling noise is in fact lower for the honest estimator.

### 3.2.3 Approach to Detect Heterogeneity

As seen above, Causal Forests automatically detect heterogeneity in treatment effects due to their tailored split criterion. However, beyond being able to predict the differences in treatment effects, two fundamental considerations are important. Firstly, we want to know whether the detected heterogeneity is of statistical significance. Secondly, heterogeneity is only of value if it allows the identification of subgroups with distinct treatment effects based on covariates. The following section describes different ways to investigating the estimates of a Causal Forest. These methods are demonstrated in Section 6.

**Intensity and Significance of Heterogeneity.** One approach to get an overview of the detected heterogeneity is to plot the variation of the predicted effects. This way, we can analyze how strongly the effects vary and compare this variation to the standard error of the average treatment effect estimate.

In the `grf` package, Athey et al. implemented a calibration test to assess a Causal Forests quality. This procedure is motivated by Chernozhukov et al. (2017). In this test, the omnibus hypothesis of heterogeneity is evaluated on held out data by computing the best linear fit of the target estimand using the mean forest prediction and the CATE-predictions as the sole two regressors. Thus, the results of this test are the coefficients for both regressors. Ideally, both coefficients are 1, suggesting that the mean forest prediction is correct and that the forest correctly captures heterogeneity. Further, the p-value of the second coefficient reveals the result of an omnibus test for the presence of heterogeneity, that is, whether we can reject the null hypothesis of no heterogeneity. Therefore, this calibration test can be used

to determine whether the detected heterogeneity is statistically significant.

**Subgroup Analysis.** Further, Athey and Wager (2019) suggest an approach for identifying subgroups with treatment effects below and above the median CATE estimate. By dividing the sample into two groups according to their CATE estimate and estimating the average treatment effect in each group, one can determine the magnitude of the difference between the groups and its statistical significance. Hereby, the definition of the subgroups is based on the complex underlying split history of all trees.

Additionally, we can still identify the impact of each variable and grasp the complex structure of the heterogeneity. With the help of a variable importance plot we can access the effect of the covariates. For this, the importance of a variable is calculated as the weighted sum of the times a tree split on it in each depth of the tree. A variable that the tree split on often apparently has a great impact that can be identified via visualizations or hypothesis testing.

### 3.3 Discussion on Forests

Extending Random Forests into Causal Forests is a considerable achievement for enabling to benefit from their advantages. Forest-based methods have become a popular and successful statistical estimator that is widely applied. Their success is not without reason.

Firstly, forests are able to flexibly adapt to various different functional forms. Being a non-parametric estimator, no assumptions are needed about the form of the data generating process or its density distribution. Having their origin in the area of Supervised Machine Learning, forests use data to determine relationships between covariates and the outcome. Nevertheless, by using techniques such as cross-validation or honesty, problems such as overfitting can be avoided.

This data-driven approach is particularly interesting in high-dimensional data sets. Forests select those variables for splitting that affect the outcome most strongly and ignore others. This way, they can be applied in settings in which the number of covariates is large, potentially even higher than the number of observations. This is in contrast to other methods, such as k-Nearest-Neighbors or local maximum likelihood estimation with kernel weighting functions that are prone to a strong curse of dimensionality (Hastie et al., 2008).

Even though most Machine Learning methods are criticized for their inability to interpret results, forests need not to be treated as black-box heuristics. By analyzing splits, forests can be interpreted. For example, one can identify variables that are split on most and investigate

their implications.

Finally, forests ease of use is a major plus. They can be applied off-the-shelf to a wide range of classification and regression problems. In addition, many computationally efficient implementations exist that require only little tuning.

Transferring these advantages to the setting of treatment effect estimation is of considerable interest. The extension of Random Forests into the Generalized Random Forests framework seems promising to enable exactly this.

# 4 Simulation Study

To investigate the performance and behavior of Causal Forests, I conducted a broad simulation study. The purpose of this simulation study is to understand how Causal Forests work and on how to make best use of them. My main interest is not in showing their competitive performance in comparison to other methods. For further information on this, I refer to Knaus et al. (2018) and Keele and Small (2018).

In the first part of this Section, I assess how well Causal Forests perform in different scenarios. In particular, I focus on the challenges mentioned in Section 2: Adequately detecting treatment effect heterogeneity, adjusting for confounding variables and a combination of both. Performance is evaluated using the MSE, absolute bias, variance estimates and coverage rates on a test sample. These results are averaged over 100 repetitions. Beyond that, I always add an example of a single forest by visualizing predictions to allow a better grasp of the numbers.

In the second part, I further explore the concept of honesty to understand its implications. For this purpose, honest Causal Forests are compared with their adaptive (not honest) analog.

Before describing the simulations and presenting the results, I explain how Causal Forests are implemented and describe the data generating processes I have chosen.

## 4.1 Implementation of Causal Forests

Causal Forests are implemented in the `R` package `grf` which is available from `CRAN`. In this package, a Causal Forest is trained based on a training sample and is used to estimate conditional average treatment effects, $\tau(x)$. These treatment effects can either be estimated on the training sample itself, using out-of-bag predictions, or can be estimated on a test sample. In addition, Causal Forests provide an estimate of the predictions variance.

By default, honest trees are grown. However, we can specify the fraction of the training sample used for determining splits or alternatively, we can completely disable honesty. Furthermore, outcomes are locally centered as described in Section 3. However, the method allows the user to provide the marginal expectations of $Y_i$ and $W_i$. By setting these expectations to zero, the centered outcomes are equivalent to the observed outcomes. This way, we can disable local centering.

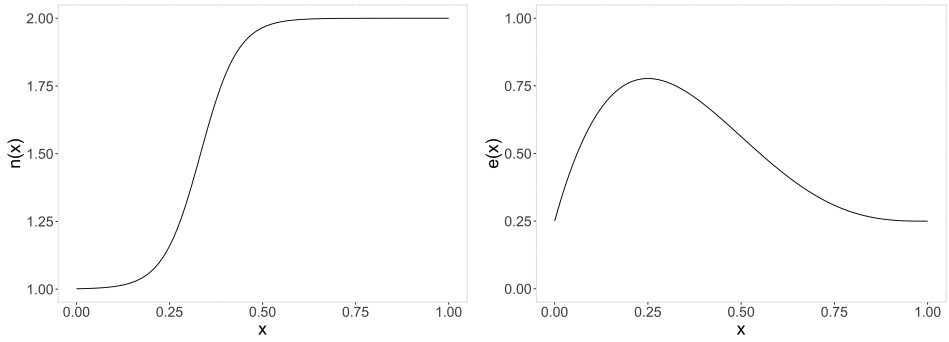
## 4.2 Data Generating Processes

Throughout this simulation study, I rely on four fundamental designs. The designs differ in the effect of the covariates. In each design, some covariates affect the treatment effect, some affect the main outcome, some affect the probability of receiving the treatment and some are just noise. Design 1 addresses the challenge of treatment effect heterogeneity, Design 2 incorporates confounders and the Design 3 combines both, heterogeneity and confounding. Finally, Design 4 incorporates treatment effect heterogeneity but focuses on adding noise. In each design, I consider two different functional forms: a simple linear and a complex non linear. The parameter $d$ denotes the dimension of the covariate space $\mathbb{X} \sim \mathcal{U}([0,1]^d)$. In each design, homoscedastic noise is added. The designs are presented in detail in their respective subsections.

The following non-linear function characterizes how a covariate enters the treatment effect function and is adopted from Wager and Athey (2018). It is defined as

$$\eta(x) = 1 + \frac{1}{1 + e^{-20(x - \frac{1}{3})}}. \tag{19}$$

Figure 2a shows a plot of this function.




**(a)** Non-linear functional form used in all designs

**(b)** Propensity Function used in Design 2 and 3

**Figure 2:** Plots of functions used in this simulation study. The left panel shows the non-linear function that is used to determine how a covariate enters the treatment effect function. The right figure shows a plot of the propensity score function used in Deisgn 2 and 3.

## 4.3 Performance Evaluation

In this section, the performance of Causal Forests is examined in three different scenarios. First, their ability to detect treatment effect heterogeneity is assessed. Second, we evaluate how well they adjust to confounding. Third, both challenges are combined.

### 4.3.1 Treatment Effect Heterogeneity

The result of the simulation study on data drawn according to Design 1 gives insight into the performance of Causal Forests in case of strong treatment effect heterogeneity.

DESIGN 1: STRONG TREATMENT EFFECT HETEROGENEITY

In this design, the treatment effect varies with covariates $X_1$ and $X_2$. In addition, the main outcome depends on $X_2$ and $X_3$. The propensity score is constant for all individuals.

| **a) Linear** | **b) Non-linear** |
|---|---|
| $\tau(x) = 2X_1 + X_2$ | $\tau(x) = \eta(X_1) \cdot \eta(X_2)$ |
| $Y = X_2 + X_3 + W_i \cdot \tau(x) + \varepsilon$ | $Y = \eta(X_2) \cdot \eta(X_3) + W_i \cdot \tau(x) + \varepsilon$ |

and

$$W_i \sim \mathcal{B}(e(x)) \text{ with } e(x) = 0.5$$

$$\varepsilon \sim \mathcal{N}(0, 0.5)$$

Thus, the ability to detect this treatment effect heterogeneity is tested.

For both, the linear and the nonlinear design, each simulation was conducted 100 times and the results were averaged. Results are evaluated in terms of MSE between the true treatment effect and the estimate of the CATE. For each simulation, I used a training size of $n = 2000$.

Since Causal Forests can be seen as a form of adaptive nearest neighbors, I compare them to k-Nearest-Neighbors (k-NN), a non-adaptive nearest neighbor method. *K-NN* estimates the conditional expectation of $Y$ in the treated and control groups separately by averaging the outcomes of the k closest training examples. Note that *close* is equally defined over the entire feature space, thus, no respect is paid to the question whether or not a feature even affects the treatment. By computing the difference between the treated and control group, the treatment effect is estimated.

Figure 3 shows the result for varying dimensions of the feature space. We see how well Causal Forests perform and how their dominance is even more pronounced as the dimension of the feature space gets large. The explanation for this is their adaptive partitioning scheme, that allows to ignore noise variables by not splitting on them. This is in contrast to k-NN,

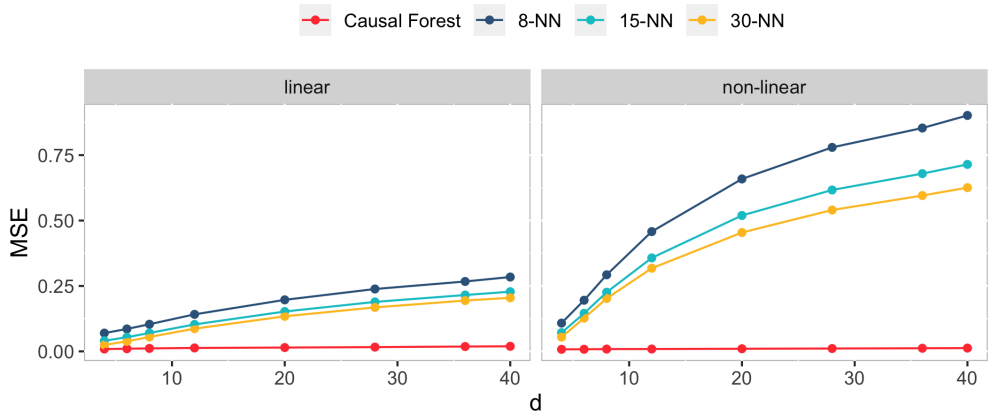



**Figure 3:** Result of simulations on data drawn according to Design 1, evaluated using the Mean-Squared-Error. The parameter $d$ denotes the dimension of the feature space.

which perform poorly as $d$ increases. A look at the variable importance plots[1] in Figure 4 confirms this explanation. Clearly, the Causal Forest identifies variables that do not affect the treatment effect. It is reassuring, that $X_3$ is considered irrelevant as it does not affect the treatment effect. Even though $X_3$ impacts the main outcome $Y$, its effect is regressed out in the local centering step.

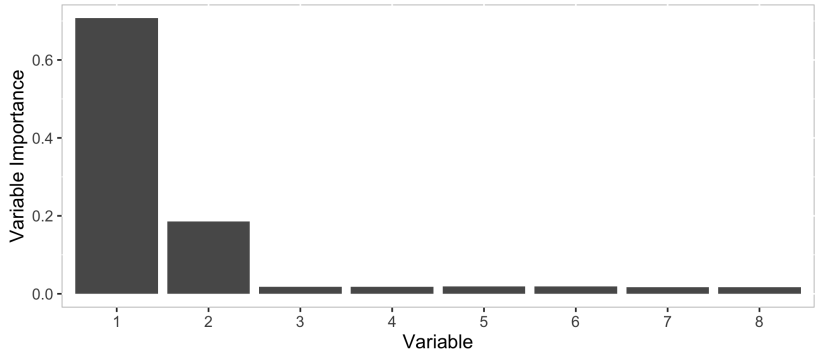
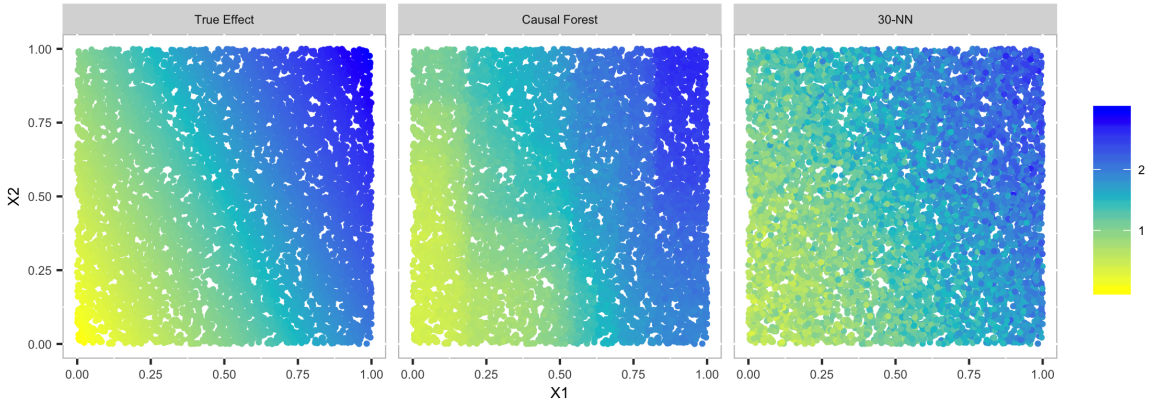


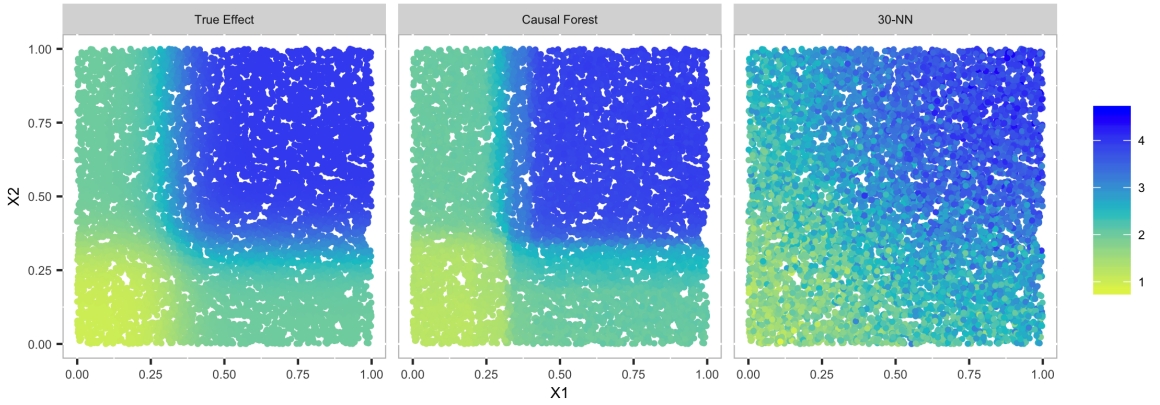

**Figure 4:** Variable Importance Plot of the Causal Forest from Simulation 1a with $d = 6$.

Taking a closer look at the specific treatment effect estimates in Figure 5, we can see how the estimates approximate the shape of the treatment effect. Although the Causal Forest estimates nearly resemble the true shape of the treatment effect, the transition boundaries are sharper. Further, the estimates are not as extreme as the true effect especially near the edges of the feature space. This is what Wager and Athey (2018) refer to as *filling the valleys and flatten the peaks*, which is typical for nearest neighbor approaches. Finally, k-NN only

[1]Variable Importance is a simple measure of the relevance of each feature. It is calculated as a weighted sum of how many times a feature was split on at each depth in a forest.




**(a)** Linear




**(b)** Non-linear

**Figure 5:** Visualization of the true treatment effect and estimates from different a Causal Forest and k-NN using data according to Design 1 with $d = 10$. The color of the points denote the magnitude of the treatment effect.

captures the general structure and is extremely grainy, indicating their poor performance seen in Figure 3.

Detailed results of the performance of Causal Forest can be seen in Table 1. MSE and absolute bias are low regardless of $n$ and $d$. However, the coverage rates for the 95% confidence intervals differ considerable from 95% and are too low. Especially as $d$ gets large, the coverage rates are as low as 69%.

| | | linear | | | non-linear | | |
|---|---|---|---|---|---|---|---|
| n | d | MSE | Bias | Coverage | MSE | Bias | Coverage |
| | 6 | 0.0366 | 0.0262 | 0.8112 | 0.0315 | 0.0273 | 0.8585 |
| 2000 | 10 | 0.0451 | 0.0306 | 0.7651 | 0.0344 | 0.0306 | 0.8216 |
| | 20 | 0.0577 | 0.0275 | 0.6945 | 0.0410 | 0.0292 | 0.7646 |
| | 6 | 0.0230 | 0.0249 | 0.8462 | 0.0201 | 0.0251 | 0.8748 |
| 4000 | 10 | 0.0281 | 0.0237 | 0.7998 | 0.0217 | 0.0239 | 0.8454 |
| | 20 | 0.0353 | 0.0234 | 0.7446 | 0.0261 | 0.0234 | 0.7850 |
| | 6 | 0.0152 | 0.0156 | 0.8707 | 0.0201 | 0.0251 | 0.8748 |
| 8000 | 10 | 0.0177 | 0.0154 | 0.8355 | 0.0217 | 0.0239 | 0.8454 |
| | 20 | 0.0208 | 0.0161 | 0.8014 | 0.0261 | 0.0234 | 0.7850 |

**Table 1:** Performance of Causal Forest on data drawn according to Design 1.

### 4.3.2 Confounding

The aim of this simulation is to evaluate the ability of Causal Forests to adjust for confounders. Therefore, the impact of local centering is assessed by comparing Causal Forests with and without this technique.

DESIGN 2: CONFOUNDING

In Design 2, there is an interaction between the main effect and the treatment assignment. This is because both, the propensity score and the main outcome, depend on the same covariate $X_1$. The true treatment effect is zero.

| **a) Linear** | **b) Non-linear** |
|---|---|
| $Y = 2X_1 + W_i \cdot \tau(x) + \varepsilon$ | $Y = \eta(X_1) + W_i \cdot \tau(x) + \varepsilon$ |

and

$$W_i \sim \mathcal{B}(e(x)) \text{ with } e(x) = \frac{1}{4}(1 + \beta_{2,4}(X_1))$$

$$\tau(x) = 0$$

$$\varepsilon \sim \mathcal{N}(0, 0.5)$$

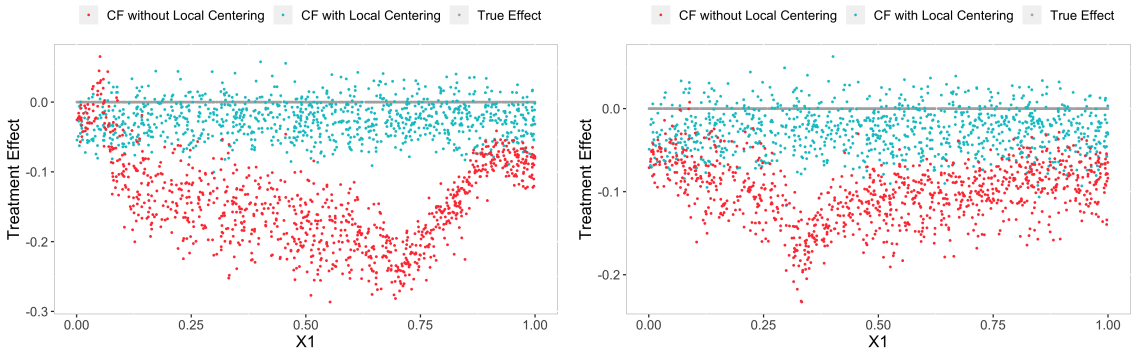
$\beta_{2,4}$ denotes the Beta distribution with shape parameters $2, 4$. Figure 2b shows a plot of the propensity score function. Put in words, units with approximately $X_1 \in [0.1, 0.5]$ are more likely to be treated than others. Note that these are fairly low values for $X_1$ and the main outcome increases with $X_1$. This implies that the average outcome of treated units is lower than for control units, even though the treatment effect is zero.

Table 2 shows how local centering improves the estimates in the presence of confounding. Causal Forests with Local Centering yield almost nominal coverage. Without local centering, the estimates are considerably biased, especially as $d$ increases. This also leads to the poor coverage rates, since confidence intervals are not centered anymore.

| | with local centering | | | without local centering | | |
|---|---|---|---|---|---|---|
| d | MSE | Bias | Coverage | MSE | Bias | Coverage |
| 4 | 0.003 | 0.022 | 0.934 | 0.029 | 0.144 | 0.413 |
| 6 | 0.002 | 0.022 | 0.950 | 0.039 | 0.171 | 0.290 |
| 8 | 0.002 | 0.020 | 0.963 | 0.045 | 0.190 | 0.244 |
| 10 | 0.002 | 0.021 | 0.967 | 0.052 | 0.207 | 0.199 |
| 12 | 0.002 | 0.020 | 0.971 | 0.055 | 0.213 | 0.187 |
| 14 | 0.001 | 0.018 | 0.978 | 0.062 | 0.227 | 0.164 |
| 16 | 0.001 | 0.019 | 0.978 | 0.062 | 0.228 | 0.157 |

**Table 2:** Comparison of Causal Forests with and without local centering in the presence of confounding, data drawn according to Design 2a.

Inspecting the specific estimates in Figure 6, one can clearly see this bias. Not surprisingly, the estimates are downwards biased due to the fact that the outcomes of the treated units are lower on average. Thus, the Causal Forest without local centering wrongly interprets the systematic differences between treated and control units as the treatment effects.




**(a)** Linear **(b)** Non-linear

**Figure 6:** Illustration of the treatment effect estimates in the presence of confounding on data drawn according to Design 2 ($X_1$ is the confounder). Causal Forests without local centering are considerably biased. Parameter values $n = 2000$, $d = 10$.

### 4.3.3 Heterogeneity and Confounding

Finally, having seen that Causal Forests can manage treatment effect heterogeneity and confounding separately, we now examine the performance in a combination of both. This case is particularly interesting since only a few traditional methods exist that address both. Table 3 outlines the results for various parameter values. Overall, the results are satisfactory. The Causal Forests is correctly adjusting for the confounding effect and detects heterogeneity.

DESIGN 3: HETEROGENEITY & CONFOUNDING

Design 3 combines both situations and contains covariates that affect the treatment effect $\{X_1, X_2\}$ as well as a confounding variable $\{X_1\}$. Moreover, $\{X_3\}$ affects the main outcome.

| **a) Linear** | **b) Non-linear** |
|---|---|
| $Y = X_1 + X_3 + W_i \cdot \tau(x) + \varepsilon$ | $Y = \eta(X_1) \cdot \eta(X_3) + W_i \cdot \tau(x) + \varepsilon$ |
| $\tau(x) = 2 \cdot (1 - X_1) + X_2$ | $\tau(x) = 2 \cdot (2 - \eta(X_1)) \cdot \eta(X_2)$ |

and

$$W_i \sim \mathcal{B}(e(x)) \text{ with } e(x) = \frac{1}{4}(1 + \beta_{2,4}(X_1))$$

$$\varepsilon \sim \mathcal{N}(0, \sigma) \text{ with } \sigma \in \{1, 1.5\}$$

To be able to make best use of Causal Forests, I analyze the implications of different parameter values in the following. As expected, performance in terms of MSE and bias improves as the training sample size ($n$) increases. The same applies to coverage rates, yet results are not as desired even for large $n$. To attain better confidence intervals, Athey et al. recommend increasing the number of trees in each forest. However, in all my simulations I find that doing so does not result in a substantial improvement of coverage rates.

Increasing the noise variance ($\sigma$) increases MSE, bias and the standard error. The change in coverage rates is negligible and they are still not as desired, even though a noise variance of 1.5 is high compared to the ranges of possible treatment effects.

In all cases, the MSE is lower for $d = 6$ than for $d = 20$, but with $n = 10000$ this improvement is only small since both values are on an excellent level.

| | | | | num.tree = 2000 | | | | num.tree = 4000 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | n | d | noise | MSE | Bias | Std. Dev. | Coverage | MSE | Bias | Std. Dev. | Coverage |
| non-linear | 2000 | 6 | 1 | 0.0465 | 0.0416 | 0.1701 | 0.8906 | 0.0463 | 0.0412 | 0.1692 | 0.8976 |
| | 2000 | 20 | 1 | 0.0583 | 0.0435 | 0.1674 | 0.8147 | 0.0579 | 0.0437 | 0.1649 | 0.8092 |
| | 5000 | 6 | 1 | 0.0264 | 0.0243 | 0.1364 | 0.9014 | 0.0259 | 0.0244 | 0.1335 | 0.9022 |
| | 5000 | 20 | 1 | 0.0314 | 0.0233 | 0.1309 | 0.8727 | 0.0309 | 0.0231 | 0.1255 | 0.8651 |
| | 10000 | 6 | 1 | 0.0187 | 0.0165 | 0.1204 | 0.9183 | 0.0186 | 0.0163 | 0.1161 | 0.9139 |
| | 10000 | 20 | 1 | 0.0206 | 0.0218 | 0.1150 | 0.9053 | 0.0205 | 0.0216 | 0.1080 | 0.8927 |
| | 2000 | 6 | 1.5 | 0.0868 | 0.0630 | 0.2367 | 0.8929 | 0.0861 | 0.0618 | 0.2323 | 0.8898 |
| | 2000 | 20 | 1.5 | 0.1079 | 0.0634 | 0.2281 | 0.8203 | 0.1090 | 0.0640 | 0.2214 | 0.8051 |
| | 5000 | 6 | 1.5 | 0.0486 | 0.0358 | 0.1885 | 0.8977 | 0.0474 | 0.0358 | 0.1845 | 0.8971 |
| | 5000 | 20 | 1.5 | 0.0567 | 0.0333 | 0.1806 | 0.8683 | 0.0557 | 0.0337 | 0.1714 | 0.8548 |
| | 10000 | 6 | 1.5 | 0.0312 | 0.0249 | 0.1697 | 0.9292 | 0.0312 | 0.0253 | 0.1613 | 0.9204 |
| | 10000 | 20 | 1.5 | 0.0327 | 0.0307 | 0.1607 | 0.9193 | 0.0327 | 0.0311 | 0.1495 | 0.9061 |
| linear | 2000 | 6 | 1 | 0.0512 | 0.0406 | 0.1579 | 0.8033 | 0.0509 | 0.0399 | 0.1591 | 0.8128 |
| | 2000 | 20 | 1 | 0.0781 | 0.0410 | 0.1497 | 0.6816 | 0.0785 | 0.0397 | 0.1486 | 0.6801 |
| | 5000 | 6 | 1 | 0.0285 | 0.0234 | 0.1326 | 0.8459 | 0.0285 | 0.0231 | 0.1310 | 0.8453 |
| | 5000 | 20 | 1 | 0.0473 | 0.0219 | 0.1263 | 0.7152 | 0.0472 | 0.0237 | 0.1215 | 0.7009 |
| | 10000 | 6 | 1 | 0.0183 | 0.0155 | 0.1187 | 0.8791 | 0.0182 | 0.0155 | 0.1159 | 0.8791 |
| | 10000 | 20 | 1 | 0.0260 | 0.0259 | 0.1148 | 0.8033 | 0.0258 | 0.0261 | 0.1098 | 0.7903 |
| | 2000 | 6 | 1.5 | 0.0799 | 0.0610 | 0.2098 | 0.8257 | 0.0793 | 0.0603 | 0.2087 | 0.8298 |
| | 2000 | 20 | 1.5 | 0.1197 | 0.0600 | 0.1928 | 0.6984 | 0.1193 | 0.0581 | 0.1878 | 0.6906 |
| | 5000 | 6 | 1.5 | 0.0467 | 0.0333 | 0.1762 | 0.8530 | 0.0461 | 0.0333 | 0.1720 | 0.8514 |
| | 5000 | 20 | 1.5 | 0.0761 | 0.0318 | 0.1634 | 0.7231 | 0.0758 | 0.0320 | 0.1537 | 0.6936 |
| | 10000 | 6 | 1.5 | 0.0307 | 0.0241 | 0.1623 | 0.8960 | 0.0303 | 0.0239 | 0.1548 | 0.8864 |
| | 10000 | 20 | 1.5 | 0.0443 | 0.0362 | 0.1520 | 0.8038 | 0.0437 | 0.0371 | 0.1419 | 0.7843 |

**Table 3:** Results of simulation on data drawn according to Design 3, in the presence of strong treatment effect heterogeneity and confounding. The parameter $n$ denotes the size of the training sample, $d$ denotes the dimensions of the feature space and *noise* is the variance of the homoscedastic noise that is added.

## 4.4 Concept Analysis: Honesty

To further investigate the concept of honesty, I compare honest Causal Forests with their adaptive analog. The honest estimator divides the training sample into two halves, while the adaptive version uses the entire training sample for both, tree construction and estimation. By adding independent and identically distributed random noise to the observed outcomes in the training sample, I examine the risk of modeling this noise for each estimator. Using a simple model with only one covariate affecting the treatment effect, the implications of honest estimation can be seen clearly.

DESIGN 4: NOISE

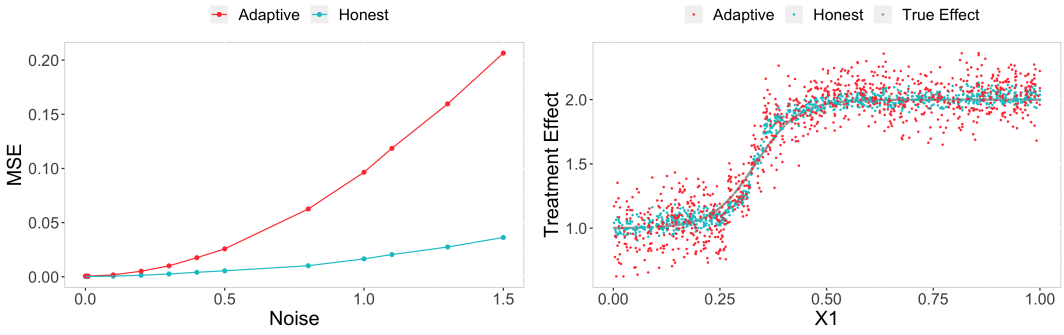
In this design, normal distributed noise with mean zero is added to the observed outcomes. The amount of this noise is enlarged by increasing its standard deviation.

$$Y = \eta(X_2) + W_i \cdot \tau(x) + \varepsilon$$

$$\tau(x) = \eta(X_1)$$

$$W_i \sim \mathcal{B}(e(x)) \text{ with } e(x) = 0.5$$

$$\varepsilon \sim \mathcal{N}(0, \sigma) \text{ with } \sigma \in [0, 2]$$




**(a)** MSE **(b)** Estimates

**Figure 7:** Demonstration of the implications of honesty. The left panel shows the MSE of a honest and an adaptive (not honest) Causal Forest on data drawn according to Design 4. On the right, the specific predictions can be seen.

Figure 7a shows, that the risk of modelling noise is in fact substantially lower for the honest

estimator. As $\sigma$ increases, the performance of the adaptive estimator declines considerably. Figure 7b illustrates how noisy its predictions become. In contrast, the honest estimator does not overfit to the noisy observed outcomes but rather captures the general structure. Although the initial motivation of Athey and Imbens (2016) behind honesty was to eliminate bias, it additionally prevents overfitting and reduces MSE.

# 5 Applying Causal Forests: Data

In the following I explore how useful Causal Forests are when analyzing an empirical data set. I revisit a field experiment from political science of Broockman and Kalla (2016), who showed the effect of 10-minute door-to-door canvassing on transphobia in Miami. In their results, they find a significant positive effect, claiming that brief but high quality conversations can durably reduce prejudice against transgender people for at least three months. Particularly interesting for my thesis, the authors did not find any heterogeneity in this treatment effect when using a simple linear model. However, treatment effect heterogeneity could be of considerable interest to better target such interventions. Therefore, I make use of a Causal Forests ability to automatically detect heterogeneity in treatment effects.

Künzel et al. (2019) analyze this data set with the same intention of finding heterogeneity and claim that the positive effect is only found among a subset of respondents. However, besides this statement they do not provide information on how to target this subgroup, which is essential in order to benefit of the detected heterogeneity. My attention will be paid to investigating whether I can confirm their findings, that is, whether I find heterogeneity in treatment effects. Further, I analyze the differences in treatment effects and aim to find subgroups with distinct effects.

## 5.1 The field experiment: Reducing Transphobia

In 2014, voters in Miami were able to decide on a law protecting transgender people from discrimination in housing, employment and public accommodations. Having seen how campaigns influence a voter's choice, the authors aim to target prejudice against transgender people. Therefore, canvassers went to voter's doorsteps and held an on average 10-minute conversation, trying to encourage subjects to actively consider the perspective of transgender people. Canvassers informed voters that they will face a decision about this issue and presented arguments on both sides. Some canvassers were themselves transgender and revealed their identity to voters during the conversations. The experiment was placebo-controlled, that is, a control group was targeted with a conversation about recycling.

To measure the social and political attitudes, the authors conducted online surveys, presented as a broad university-sponsored public opinion survey. Each survey included several unrelated questions, with only a few concerning transphobia. First, registered voters were recruited for a baseline survey (n=68378) and then, respondents (n=1825) were randomly assigned to either the treatment (n=913) or the placebo (n=912). Randomization was con-

ducted on a household level, ensuring that all individuals in the same household received the same treatment. Finally, all voters that come to their door (n=501) were recruited to follow-up online surveys at four different points in time: 3-day (n=429), 3 weeks (n=399), 6 weeks (n=401) and 3 months (n=385) after the intervention.

### 5.2 Data & Estimation Details

**Covariates.** 28 pre-treatment covariates were considered, all from the baseline survey at $t = 0$. Some describe baseline characteristics of the voters, such as their age, gender, religion and political affiliation. Others also capture their view on transgender people before the intervention and their general attitude towards gender prejudices. All covariates and their meaning are listed in Table 7 in the appendix, but those important for further analysis will be described when mentioned. Before the intervention, there were no systematic differences between the treatment and control group. On all covariates, both groups were equal on average. This was achieved by randomly assigning individuals to the treatment.

**Outcomes.** In each survey, various questions regarded the voter's attitude towards transgender, including a general feeling thermometer, questions regarding the law and regarding the attitude towards gender norms. Most outcomes are measured on a $-3$ to $+3$ Likert scale. In order to aggregate this information, the authors combine multiple items into indices, of which I consider the following three:

- Index 1: "All" *(Contains all primary outcomes, created to test the omnibus hypothesis whether the intervention had any effect)*
- Index 2: "Transgender Tolerance" *(Contains questions that capture acceptance and tolerance towards transgender)*
- Index 3: "Law" *(Contains attitudes regarding Miami's law protecting transgender people from discrimination)*

The indices are defined as the first principle component of the combination of the outcomes included in the index. The factors are re-scaled to mean 0 and standard deviation 1 to allow for a natural interpretation of the size of the effects.

**Estimation**. The authors estimate treatment effects using ordinary least squares regres-

sion (OLS) with cluster-robust standard errors, clustering on household[2].

In a pre-analysis-plan (PAP), they specify theoretical predictions about treatment effect heterogeneity. In particular, they propose two hypotheses:

- Democrats will be more treatment responsive than Republicans.
- Subjects higher on the baseline support scale will be more treatment responsive as a result of being more open to outgroup rights in general.

They computed a residualized index and then tested whether the treatment effect on this index was larger for any subgroup. This way, they ensure that the outcomes do not reflect baseline differences among the subgroups. In addition, they implemented a procedure that relies on the Lasso, a variable selection algorithm, to search for any robust heterogeneous treatment effects. However, they found none.

### 5.3 Original Findings

The authors, Broockman and Kalla, claim the intervention was broadly successful at increasing acceptance against transgender. Before canvassing, both groups scored similarly on the indices. Afterwards, the treatment group was considerably more accepting. These effect were considered to be long-lasting as the treatment group remained more accepting in the follow-up surveys. Further, the conversations were broadly effective, for both democrats and republicans as well as for subjects that were more and less supportive than the average in the baseline survey. Detailed results of their ATE-estimates can be found along with my results in Section 6. No evidence was found for heterogeneity in treatment effects.

[2]The authors also adjust for measurement errors in the treatment indicator due to inadvertently delivering the treatment to some individuals in the placebo group. This way, they estimate the Complier Average Causal Effect, which slightly increases point estimates. However, since the information provided on this in their supplemental materials is insufficient, I disregard this adjustment and estimate the ATE on the intend-to-treat. In comparison to the Complier Average Causal Effect, this evaluates the overall effect in the real world and does not adjust for the fact that some voters did not comply.

# 6 Results

Using the same covariates and outcome indices, I explore the data using a Causal Forest. Starting with estimating the ATE on all indices for all points in time, I move on to my main question of interest: Can a Causal Forest find any significant treatment effect heterogeneities based on observable characteristics? I present my results on this questions by first showing my analysis for Index 2 evaluated at the three-day survey in order to better illustrate my approach and findings. Afterwards, I rerun this procedure for all other surveys and indices. Finally, I compare my results to those of Künzel et al. (2019).

## 6.1 Average Treatment Effects

Table 4 shows the estimates of the average treatment effect on all indices for all four points in time. I compare the authors' estimates to an OLS-regression that I conducted as well as to the Causal Forest estimates. For all estimators, I used cluster-robust standard errors. The results I obtained are generally lower than what the authors report, sometimes even lower than within a range of one standard deviation. One reason for this is that I do not adjust for measurement errors in the treatment assignment. However, adjusting for this error still leaves the estimates on a lower level than reported by the authors. Further, the Causal Forest estimates are slightly lower than those of an OLS regression. Despite this difference, all three estimators suggest the intervention was successful, especially regarding the first to indices. The highest effect is found on Index 2, which excludes the law items. This is in line with the authors results, who report the greatest effect on the transgender tolerance score.

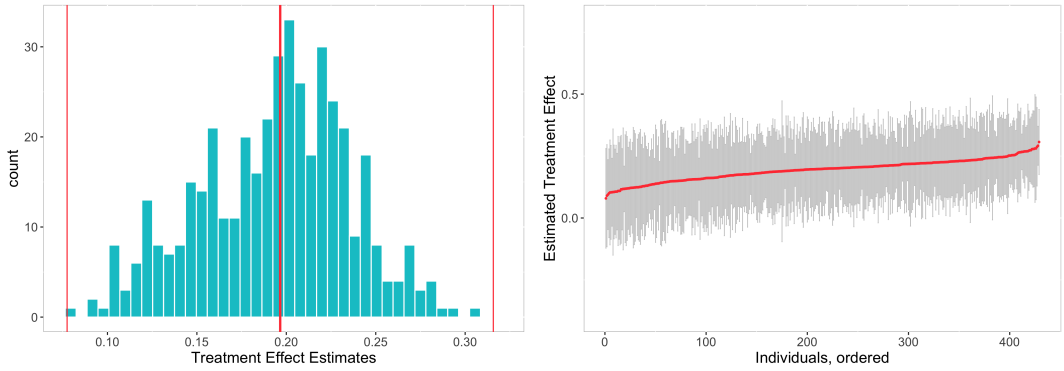
| | | t = 1 | | t = 2 | | t = 3 | | t = 4 | |
|---|---|---|---|---|---|---|---|---|---|
| | Estimator | Estimate | Std. Error | Estimate | Std. Error | Estimate | Std. Error | Estimate | Std. Error |
| Index 1 | Authors | 0.225 | 0.072 | 0.185 | 0.076 | 0.295 | 0.076 | 0.257 | 0.075 |
| | OLS | 0.168 | 0.054 | 0.136 | 0.056 | 0.219 | 0.057 | 0.193 | 0.056 |
| | Causal Forest | 0.146 | 0.061 | 0.105 | 0.061 | 0.203 | 0.062 | 0.181 | 0.060 |
| Index 2 | Authors | 0.292 | 0.072 | 0.238 | 0.079 | 0.352 | 0.074 | 0.345 | 0.078 |
| | OLS | 0.217 | 0.054 | 0.175 | 0.058 | 0.261 | 0.055 | 0.259 | 0.058 |
| | Causal Forest | 0.197 | 0.061 | 0.145 | 0.063 | 0.242 | 0.060 | 0.237 | 0.060 |
| Index 3 | Authors | 0.002 | 0.135 | 0.041 | 0.141 | 0.362 | 0.164 | 0.295 | 0.152 |
| | OLS | 0.001 | 0.068 | 0.021 | 0.069 | 0.157 | 0.071 | 0.128 | 0.066 |
| | Causal Forest | 0.009 | 0.073 | 0.007 | 0.074 | 0.151 | 0.078 | 0.125 | 0.073 |

**Table 4:** Estimates of the average treatment effect of the intervention, comparing the results the authors report with my results using OLS and Causal Forest.

## 6.2 Treatment Effect Heterogeneity: A Demonstration

Looking beyond the average treatment effect, I now investigate the differences in treatment effects. I demonstrate my approach for finding heterogeneity in detail using the outcome Index 2 evaluated at the three-day survey. I conduct the steps described in Section 3.2.3 and present my results.

A first insight into the heterogeneity of treatment effects is gained by looking at the individual predictions and their variation. Instead of estimating the ATE as in Table 4, I now compute treatment effect estimates using out-of-bag-predictions for each single individual. Figure 8a shows a histogram of out-of-bag treatment effect estimates along with the 95%-confidence interval of the ATE estimate. Clearly, the estimates show some variation around the ATE, but almost all estimates are within the confidence interval of the ATE estimate, suggesting that the variation is rather small. This can be seen even more explicitly in Figure 8b, which shows the treatment effect estimates, ordered by their magnitude along with their confidence intervals. The variation in treatment effects apparently does not exceed the uncertainty in each estimate.




**(a)** Histogram of Treatment Effects **(b)** Confidence Intervals, Ordered

**Figure 8:** The left panel shows a histogram of Treatment Effect Estimates using out-of-bag predictions for the Transgender Tolerance Index evaluated at the three-day-survey, along with the ATE estimate and the boundaries of its 95%-Confidence Interval. The panel on the right shows the 95%-Confidence Intervals for the treatment effect estimates, in ascending order.

Running a calibration test that evaluates the quality of the random forest estimates confirms this belief. A coefficient of 1 suggests the prediction is correct, which is given for `mean.forest.prediction`. The negative coefficient for `differential.forest.prediction` shows that the forest has not found adequate heterogeneity and its large p-value indicates

that we cannot reject the null hypothesis of no heterogeneity. The following excerpt shows the result.

```
Best linear fit using forest predictions (on held-out data)
as well as the mean forest prediction as regressors, along
with one-sided heteroskedasticity-robust (HC3) SEs:

                               Estimate Std. Error t value    Pr(>t)
mean.forest.prediction          1.01138    0.30833  3.2802 0.0005611 ***
differential.forest.prediction -2.74282    1.39367 -1.9681 0.9751461
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

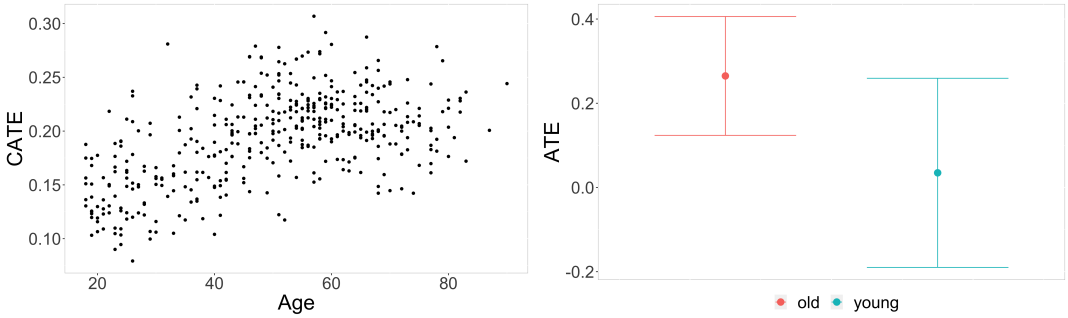
We obtain the same result when grouping observations according to whether their out-of-bag treatment effect estimate is above or below the median and estimating the ATE in each group. I expected a small difference between the estimates that is not statistically significant. Surprisingly, the opposite occurred, that is, the estimate for the *below-median* group is much larger than for the *above-median*. I assume this is due to the notably insignificant heterogeneity as seen in the calibration test.

Even though the variation in treatment effect estimates is small and the omnibus tests did not find strong evidence of heterogeneity, we might still be interested in what is causing these differences. As Athey and Wager show, such analysis can still provide us with valuable insight. The variable importance of the Causal Forest suggests taking a closer look at the covariates shown in Table 5. The forest spent most splits on the variable *age*.

| | Age | Therm. Gay | Therm. Obama | Therm. Trans. |
|---|---|---|---|---|
| Variable Importance | 0.188 | 0.087 | 0.081 | 0.074 |

**Table 5:** Variable Importance for top 4 covariates. *Therm.* refers to the feeling thermometer towards 1) gay men, 2) Obama 3) Transgender, all values come from the baseline survey.

If we manually analyze the impact of the covariate *age*, we see a clear trend, see Figure 9a. The positive effect of the intervention is much more pronounced for older voters. To determine whether this difference is of statistical significance, I divide the sample into two groups, according to their age and estimate the average treatment effect in each. Note that the boundary between both groups (40 years) is somewhat heuristically motivated. The estimates show a notable difference. Even though this gives a tiny qualitative insight into the structure of the heterogeneity, the result of running a hypothesis test shows that this difference is not statistically significant, which is reassuring.




**(a)** CATE Estimates **(b)** ATE Estimates

**Figure 9:** Impact of covariate *age* on CATE Estimates

Proceeding like this with the other seemingly relevant covariates confirms that there is no significant heterogeneity in treatment effects based on covariates. As expected, none of the items the authors suspected to cause heterogeneity in their PAP appear to have an impact, which is in line with their findings.

### 6.3 Heterogeneity Analysis on all Indices

I repeated this analysis on the other indices to all four points in time. In none of these cases, the CATE estimates differ considerably from the ATE. A summary of my findings regarding the heterogeneity in treatment effects can be found in Table 6 and Figure 10. These numbers document how small the variation around the ATE is, suggesting that there is only little heterogeneity.

Running the calibration test on each outcome confirms this. In all cases, the coefficient for `mean.forest.prediction` is almost 1 while for `differential.forest.prediction` it is strongly negative. These results can be seen in Table 8 in the appendix.

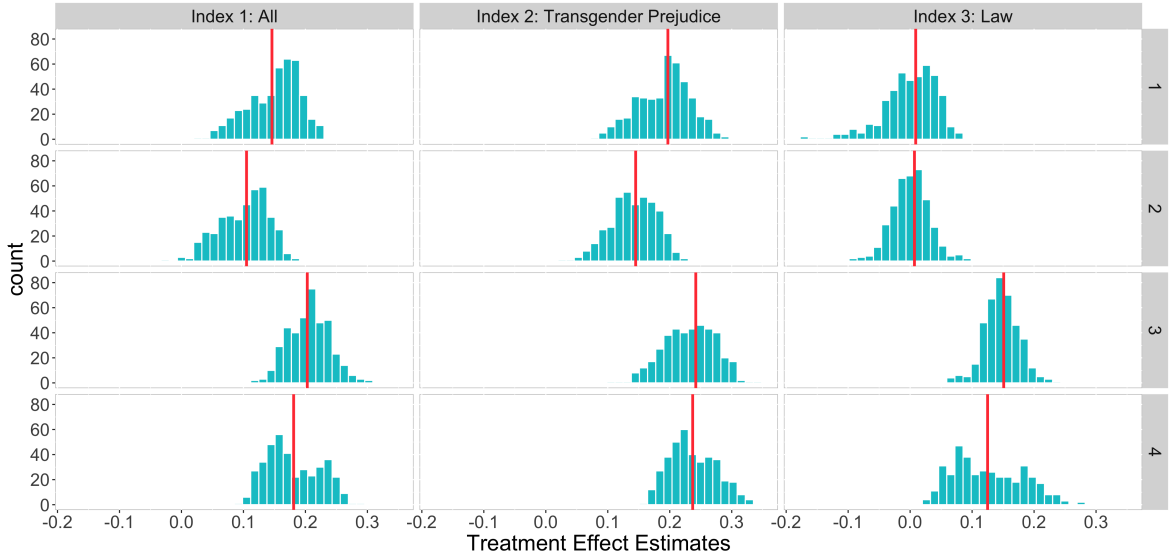



**Figure 10:** Histograms of treatment effect estimates using out-of-bag prediction, showing the variation of CATE, for the for all three indices and all four points in time (t=1,2,3,4). The vertical red line depicts the ATE estimate.

| | t | Mean | S.D. | Min | Max | S.E. |
|---|---|---|---|---|---|---|
| Index 1 | 1 | 0.150 | 0.041 | 0.027 | 0.230 | 0.088 |
| | 2 | 0.101 | 0.039 | -0.024 | 0.186 | 0.086 |
| | 3 | 0.206 | 0.033 | 0.116 | 0.298 | 0.079 |
| | 4 | 0.179 | 0.043 | 0.094 | 0.284 | 0.078 |
| Index 2 | 1 | 0.191 | 0.042 | 0.079 | 0.293 | 0.093 |
| | 2 | 0.141 | 0.036 | 0.028 | 0.223 | 0.084 |
| | 3 | 0.232 | 0.041 | 0.102 | 0.337 | 0.086 |
| | 4 | 0.237 | 0.037 | 0.162 | 0.330 | 0.080 |
| Index 3 | 1 | 0.000 | 0.043 | -0.175 | 0.081 | 0.097 |
| | 2 | 0.001 | 0.029 | -0.091 | 0.098 | 0.106 |
| | 3 | 0.147 | 0.027 | 0.062 | 0.236 | 0.096 |
| | 4 | 0.126 | 0.056 | 0.028 | 0.281 | 0.093 |

**Table 6:** Descriptive Statistics on Treatment Effect Estimates using out-of-bag prediction on all three indices at all four points in time. *S.D.* denotes the Standard Deviation in the estimates and *S.E.* denotes the mean of the Standard Errors of each estimate.

## 6.4 Comparison to X-Learner

In Künzel et al. (2019), the same data set is analyzed with focus on finding heterogeneity. They propose the *X-Learner*, a metaalgorithm that builds on base algorithms such as Random Forests and provides a framework to estimate CATEs. Their algorithm is specifically

designed for unbalanced or small data sets, because it exploits information from the treatment and control group separately. They apply their algorithm on two data sets, of which one is the Transgender-Prejudice study of Broockman and Kalla (2016), with the aim to find heterogeneity that has not been detected before.

Künzel et al. claim, they find that *"there is strong evidence that the positive effect the authors report is only found among a subset of respondents that can be targeted based on observable characteristics."* Further, they show the histogram of CATE Estimates for Index 2 evaluated at the three-day survey, but do not provide any other insight into their findings. Particularly, they do not show how to target the subset with the significant positive effect.

Implementing the X-Learner and comparing results with those of the Causal Forest reveals small differences in the estimates. Indeed, the variation in the X-Learner's predictions is a little larger, though the center of the estimates is still at the ATE, see Figure 11.

Certainly, I can confirm that when estimating the treatment effect for every individual, the effect is not significantly positive for all voters. This can be seen in the CATE estimates in 8b, that show that the confidence interval reaches below zero for some estimates. However, I can not confirm that the detected heterogeneity is statistically significant nor that these subsets can be targeted based on covariates. Since our aim is to better target the intervention based on the voter's baseline characteristics, heterogeneity is only of value if we can identify subgroups with distinct effects, which in this case is not given.

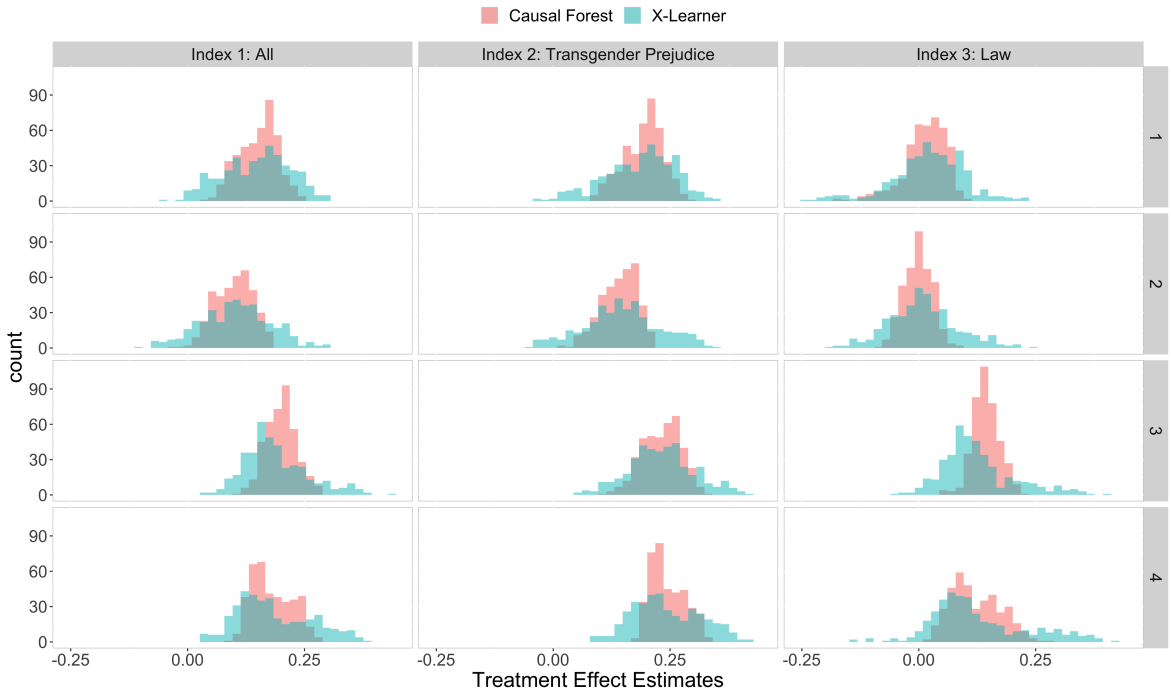



**Figure 11:** Comparison of Histograms of CATE Estimates, using the X-Learner of Künzel et al. and Causal Forest.

## 6.5 Discussion

Overall, my results confirm the findings of Broockman and Kalla (2016). Although the treatment effect estimates show some variation, neither of it is statistically significant nor can it be completely associated with variation in covariates. Thus, the intervention was broadly successful in reducing prejudice against transgender.

Concluding the practicability of Causal Forests, this analysis shows their ease of use and interpretability. Analyzing a data set with Causal Forests enables to automatically detect possible heterogeneity while still providing deep insights into its underlying structure. With the calibration test and the procedure to group observations according to the magnitude of their treatment effect estimate, two straightforward and easy-to-implement approaches exist that test the omnibus hypothesis of treatment effect heterogeneity. Especially the calibration test procedure provided by the `grf` package proves efficient and time-saving when there are many outputs to be analyzed. In addition, with the help of a variable importance plot, one can explore the impact of the most important covariates and identify subgroups with distinct treatment effects.

# 7 Conclusion

With Causal Forests, a powerful yet easy to use method for estimating heterogeneous treatment effects exists. In this thesis, I have provided intuition on the underlying conceptual ideas, have evaluated the performance, examined the concepts of honesty and local centering and demonstrated its application to a field experiment in political science. Returning to the requirements defined in the beginning of this thesis, Causal Forests successfully meet almost all of them with only small shortcomings regarding coverage rates.

Beyond estimating the average effect of a treatment, Causal Forests give insight into understanding how the treatment effect varies among subgroups. By explicitly incorporating treatment effect heterogeneity in the split criterion of trees, they automatically detect differences based on covariates. These heterogeneities can be interpreted to identify subgroups with distinct effects. Different practical approaches to inspect the detected differences exist; some perform a general analysis of whether the heterogeneity is statistically significant and others allow a manual exploration of the covariates impact. These approaches prove especially convenient for high-dimensional data sets or multiple outcomes.

Further, with the technique of local centering, Causal Forests adequately adjust for confounding effects, as the results of my simulation study show. Therefore, they are not only applicable to randomized control trials, but also to observational studies.

Since treatment effect settings often require rigorous statistical inference, confidence intervals play an essential role. Even though Athey et al. (2019) show they are theoretically valid, the results of my simulations reveal that there is potential for improvement. Coverage rates for 95% Confidence Intervals go as low as 69% in the presence of strong heterogeneity and confounding. However, exploring heterogeneity still gives valuable qualitative insights that can be used as a starting point for further research.

Everything considered, Causal Forests illustrate the value that Machine Learning can add to Causal Inference. Due to their data-driven process to determine which covariates affect the treatment effect, they handle high-dimensional data sets where the number of covariates is large and can flexibly adapt to different functional forms. A major challenge of letting data itself discover relationships is the risk of modeling spurious correlations. However, by relying on the concept of honesty, Causal Forests effectively prevent overfitting to noise, as shown in my simulations. Thus, the introduction of Causal Forests opens up vast possibilities for researchers investigating how the effect of a treatment varies among subgroups.

## References


Ascarza, E. (2018): "Retention Futility: Targeting High-Risk Customers Might be Ineffective," *Journal of Marketing Research*, 55, 80–98.

Athey, S. and G. Imbens (2016): "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences of the United States of America*, 113, 7353–7360.

Athey, S. and G. W. Imbens (2015): "Machine Learning Methods for Estimating Heterogeneous Causal Effects," .

Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized Random Forests," *The Annals of Statistics*, 1148–1178.

Athey, S. and S. Wager (2019): "Estimating Treatment Effects with Causal Forests: An Application," .

Biau, G. (2012): "Analysis of a Random Forest Model," *Journal of Machine Learning Research*, 13, 1063–1095.

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

Brand, J. E. and Y. Xie (2010): "Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education," *American sociological review*, 75, 273–302.

Breiman, L. (2001): "Random Forests," *Machine Learning*, 5–23.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984): *Classification and regression trees*, Boca Raton: Chapman & Hall, repr ed.

Broockman, D. and J. Kalla (2016): "Durably reducing transphobia: A field experiment on door-to-door canvassing," *Science*, 352, 220–224.

Bühlmann, P. and B. Yu (2002): "Analyzing Bagging," *The Annals of Statistics*, 927–961.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2017): "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," .

Denil, M., D. Matheson, and N. de Freitas (2014): "Narrowing the Gap: Random Forests In Theory and In Practice," in *Proceedings of the 31st International Conference on Machine Learning*, ed. by E. P. Xing and T. Jebara, Bejing, China: PMLR, vol. 32 of *Proceedings of Machine Learning Research*, 665–673.

Hastie, T., R. Tibshirani, and J. Friedman (2008): *The Elements of Statistical Learning*, Springer.

Hill, J. L. (2011): "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240.

Holland, P. W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 945–960.

Keele, L. and D. Small (2018): "Comparing Covariate Prioritization via Matching to Machine Learning Methods for Causal Inference using Five Empirical Applications," .

Knaus, M. C., M. Lechner, and A. Strittmatter (2018): "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence," .

Kravitz, R. L., N. Duan, and J. Braslow (2004): "Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages," *The Milbank Quarterly*, 82, 661–687.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019): "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences of the United States of America*, 116, 4156–4165.

Obermeyer, Z. and E. J. Emanuel (2016): "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine," *The New England journal of medicine*, 375, 1216–1219.

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

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 688–701.

——— (1980): "Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment," *Journal of the American Statistical Association*, 75, 591–593.

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.

Wager, S. and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242.

Wager, S., T. Hastie, and B. Efron (2014): "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife," *Journal of Machine Learning Research*, 15, 1625–1651.

Zeileis, A., T. Hothorn, and K. Hornik (2008): "Model-based Recursive Partitioning," *Journal of Computational and Graphical Statistics*, 17, 492–514.

# A Tables

**Table 7:** Baseline Covariates of voters in the field experiment regarding transgender prejudice of Broockman and Kalla (2016). Covariates starting with *vf* are from the voter file and covariates ending with *t0* are from the baseline survey.

| Name | Description |
|---|---|
| miami_trans_law_t0 | "Miami-Dade county recently passed a law that prohibits discrimination in housing, employment and public accommodations based on gender identity and expression, a category that includes transgender men and women. Do you favor or oppose this new law?" |
| miami_trans_law2_t0 | "Some people say it's important to protect transgender people from discrim- ination in housing and employment. Other people have concerns about society becoming too accepting of transgender people, and do not want transgender people included in our non-discrimination law. What do you think? Do you agree or disagree that Miami law should protect transgender people from discrimination?" |
| therm_trans_t0 | Pre-sruvey, feeling thermometer towards transgender people |
| gender_norms_sexchange_t0 | "I would support a friend choosing to have a sex change. Do you agree or disagree?" |
| gender_norms_moral_t0 | "It is morally wrong for a man to present himself as a woman in public. Do you agree or disagree?" |
| gender_norms_abnormal_t0 | "A man who identifies as a woman is psychologically abnormal. Do you agree or disagree?" |

| | |
|---|---|
| ssm_t0 | Pre-survey, same sex marriage support |
| therm_obama_t0 | Pre-sruvey, feeling thermometer towards Obama |
| therm_gay_t0 | Pre-sruvey, feeling thermometer towards gay men |
| vf_democrat | Voter File Democrat |
| ideology_t0 | Pre-survey, political ideology |
| religious_t0 | Pre-survey, religiousity |
| exposure_gay_t0 | Pre-survey, knows gay people |
| exposure_trans_t0 | Pre-survey, knows transgender people |
| pid_t0 | Pre-survey, partisanship |
| sdo_scale | Pre-survey, social dominance orientation scale |
| gender_norm_daughter_t0 | "Parents usually maintain stricter control over their daughters than their sons, and they should. Do you agree or disagree?" |
| gender_norm_looks_t0 | "To keep children from being confused, it's better when men look and act like men, and women look and act like women. Do you agree or disagree?" |
| gender_norm_rights_t0 | "Men and women should have equal rights, but men and women are not the same; it's normal for men to act like men, and women to act like women. Do you agree or disagee?" |
| therm_afams_t0 | Pre-sruvey, feeling thermometer towards african-american |
| vf_female | Voter File Female |
| vf_hispanic | Voter File Hispanic |
| vf_black | Voter File african-american |
| vf_age | Voter File Age |
| survey_language_es | Survey conducted in spanish |
| cluster_level_t0_scale_mean | Scale used for blocking at household level |

| Index | t | *mean forest prediction* | *differential forest prediction* |
|---|---|---|---|
| Index 1 | 1 | 1.015 | -2.965 |
| | 2 | 1.026 | -3.543 |
| | 3 | 0.985 | -6.139 |
| | 4 | 1.009 | -2.723 |
| Index 2 | 1 | 0.979 | -2.756 |
| | 2 | 0.987 | -3.302 |
| | 3 | 0.997 | -3.540 |
| | 4 | 0.999 | -2.501 |
| Index 3 | 1 | 1.284 | -6.204 |
| | 2 | 1.719 | -13.826 |
| | 3 | 0.997 | -16.299 |
| | 4 | 0.954 | -3.191 |

**Table 8:** Result of calibration test on each outcome index for $t = 1, 2, 3, 4$. The coefficients for *mean forest prediction* of almost 1 suggests the estimate of the average treatment effect is correct, while the negative coefficient for *differential forest prediction* indicates that the forest did not adequately detect heterogeneity.

## Declaration of Authorship

I hereby confirm that I have authored this Bachelor's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Karlsruhe, May 28, 2020

Frederike Lübeck