2018

Quantifying and Explaining Causal Effects of World Bank Aid Projects

Jianing Simulation Zhao

College of William and Mary - Arts & Sciences, zhaojianing2011@gmail.com

Follow this and additional works at: https://scholarworks.wm.edu/etd

Part of the Computer Sciences Commons

Recommended Citation
http://dx.doi.org/10.21220/s2-cjp6-5g19

This Dissertation is brought to you for free and open access by the Theses, Dissertations, & Master Projects at W&M ScholarWorks. It has been accepted for inclusion in Dissertations, Theses, and Masters Projects by an authorized administrator of W&M ScholarWorks. For more information, please contact scholarworks@wm.edu.
Quantifying and Explaining Causal Effects of World Bank Aid Projects

Jianing Zhao
Williamsburg, VA

Bachelor of Science, China Agriculture University, 2009
Master of Science, China Agriculture University, 2011

A Dissertation presented to the Graduate Faculty
of The College of William & Mary in Candidacy for the Degree of
Doctor of Philosophy

Department of Computer Science

College of William & Mary
August, 2018
This Dissertation is submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Jianing Zhao

Approved by the Committee, May 2018

Committee Chair
Associate Professor Peter Kemper, Computer Science
College of William & Mary

Associate Professor Gang Zhou, Computer Science
College of William & Mary

Assistant Professor Xu Liu, Computer Science
College of William & Mary

Assistant Professor Zhenming Liu, Computer Science
College of William & Mary

Assistant Professor Daniel Runfola, Applied Science
College of William & Mary
ABSTRACT

In recent years, machine learning methods such as deep learning have enabled us to predict with good precision using large training data. However, for many problems, we care more about causality than prediction. For example, instead of knowing that smoking is statistically associated with lung cancer, we are more interested in knowing that smoking is the cause of lung cancer. With causality, we can understand how the world progresses and how impacts are made on an outcome by influencing the cause.

This thesis explores how to quantify the causal effects of a treatment on an observable outcome in the presence of heterogeneity. We focus on investigating the causal impacts that World Bank projects have on environmental changes. This high-dimensional World Bank data set includes covariates from various sources and of different types, including time series data, such as the Normalized Difference Vegetation Index (NDVI) values, temperature and precipitation, spatial data such as longitude and latitude, and many other features such as distance to roads and rivers.

We estimate the heterogeneous causal effect of World Bank projects on the change of NDVI values. Based on analysis techniques using causal trees and causal forests proposed by Athey, we described the challenges we met and lessons we learned when applying these two methods to an actual World Bank data set. We show our observations of the heterogeneous causal effect of the World Bank projects on the change of environment. As we do not have the ground truth for the World Bank data set, we validate the results using synthetic data for simulation studies. The synthetic data is sampled from distributions fitted with the World Bank data set. We compared the results among various causal inference methods and observed that feature scaling is very important to generating meaningful data and results. In addition, we investigate the performance of the causal forest approach with various parameters such as leaf size, number of confounders, and data size.

A causal forest is a black-box model. The results obtained from it cannot be easily interpreted. We derived a refined method to compute a linear regression model for an individual project as a local approximation of the global causal forest model. By taking advantage of the tree structure, we select neighbors of the project to be explained. The weights are assigned to the neighbors according to dynamic distance metrics. We can learn a linear regression model with the neighbors and interpret the results with the help of the learned linear regression model.

In summary, World Bank projects have small impacts on the change to the environment, and the result of an individual project can be interpreted using a linear regression model learned from its vicinity in a global causal forest model.
TABLE OF CONTENTS

Acknowledgments v

Dedications vi

List of Tables vii

List of Figures ix

Chapter 1. Introduction 1
1.1 Correlation and Causality 1
1.2 Causal Inference in Observational Studies 3
   1.2.1 Potential Outcome Framework (Rubin-Neyman Causal Model) 3
   1.2.2 Common Assumptions in Causal Inference 4
      1.2.2.1 Ignorability or Absence of Unobserved Confounders 4
      1.2.2.2 Common Support or Overlap 5
      1.2.2.3 Stable Unit Treatment Value Assumption (SUTVA) 6
1.3 Heterogeneity in Causal Effects 6
   1.3.1 Decision Tree 7
   1.3.2 Causal Tree 8
1.4 Other Causal Inference Approaches 9
   1.4.1 Judea Pearl’s Causal Model 9
1.4.2 Granger Causality 10

1.5 Interpretable Model 11

1.6 Overview 11

Chapter 2. World Bank Data 12

2.1 Vegetation 15

2.2 Geographical Locations 18

2.3 Distance to Rivers 20

2.4 Distance to Roads 22

2.5 Accessibility to Urban Areas 25

2.6 Population Density 26

2.7 Slope and Elevation 29

2.8 Temperature 31

2.9 Precipitation 35

2.10 Project Starting Year 39

2.11 Project Funding 39

2.12 Data Characteristics 41

2.13 Data Interpretation 44

Chapter 3. Heterogeneous Causal Effect of World Bank Financed Projects 48

3.1 Introduction 49

3.2 Methodology 52

3.2.1 Causal Effects 52

3.2.2 Causal Tree 53

3.2.3 Causal Forest 54

3.3 Experiments and Results 54
3.3.1 Experimental Results for Simulated Data 55
3.3.2 Results for World Bank Data 57
3.4 Discussion and Conclusions 60

Chapter 4. Simulation Study in Quantifying Heterogeneous Causal Effects 62
4.1 Introduction 63
4.2 Causal Inference Techniques 66
  4.2.1 The Rubin Causal Model and Conditional Average Treatment Effects 66
  4.2.2 Regression Tree for Causal Inference 67
  4.2.3 Transformed Outcome Tree 68
  4.2.4 Causal Tree 68
  4.2.5 Random Forest 69
4.3 Simulation Model 70
  4.3.1 Configuring the Simulation Model to Approximate the Real Data Set 72
4.4 Experiments and Results 73
  4.4.1 Accuracy of Causal Effect Estimate 74
  4.4.2 Convergence Rate for Increasing Data Size 80
4.5 Discussion and Conclusions 82

Chapter 5. Seeking Explanations from Causal Forests for World Bank Aid Projects 84
5.1 Introduction 85
5.2 Methods 87
  5.2.1 Causal Tree and Causal Forest 87
5.2.2 Computation of an Explanation Model 89
5.2.3 Distance Measurement 90
   5.2.3.1 A Purely Structural Notion of Distance 90
   5.2.3.2 A Purely Data-driven Notion of Distance 91
   5.2.3.3 A Combined Notion of Distance 92
5.2.4 Sampling Data for an Explanation Model 93
5.2.5 Sparse Linear Explanations 95
5.3 Causal Effect Calculation for World Bank Data 97
   5.3.1 Zhejiang Urban Environment Project 98
   5.3.2 A Municipal Infrastructure Project in Lebanon 105
5.4 Causal Effect of World Bank Projects to Night Lights 107
   5.4.1 Discussion 109
5.5 Conclusion 111

Chapter 6. Conclusion 112
ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Dr. Peter Kemper for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance and expert advice helped me in my research and writing of this thesis. Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Daniel Runfola, Dr. Gang Zhou, Dr. Xu Liu, and Dr. Zhenming Liu, for their insightful comments. In particular, I would like to thank Dr. Daniel Runfola for his valuable feedback and discussions for my research and thesis. I would especially like to thank AidData, thanks for the valuable World Bank data and financial support. I thank all the people at the Department of Computer Science of the College of William and Mary. Last but not the least, I would like to thank my parents for their love, support, and sacrifice. I can not finish the thesis without them.
This thesis is to my parents: Cuilan Liu and Xue Zhao.
LIST OF TABLES

2.1 Covariates of the data sets for World Bank projects 14
2.2 Distance to large rivers with a potential commercial use (in km) 22
2.3 Distance to rivers (in km) 22
2.4 Distance to roads (in km) 24
2.5 Travel time (in minutes) 26
2.6 Population density in different years (in persons per square km) 28
2.7 Slope (in degree) 29
2.8 Elevation (in m) 29
2.9 Statistics for the intercept of a fitted linear regression for average, min, and max temperature time series data 32
2.10 Statistics for the slope of a fitted linear regression for average, min, and max temperature time series data 32
2.11 Statistics for the intercept of a fitted linear regression for average, min, and max precipitation time series data 36
2.12 Statistics for the slope of a fitted linear regression for average, min, and max precipitation time series data 36
2.13 Statistics on the amounts of funding of World Bank projects (in US$) 40
2.14 Moments 46
2.15 Covariate distribution fitting 47
3.1 Mean square error (forest has 1000 trees, feature ratio = 0.8) 56

4.1 MSE results for different configurations of the splitting rule. 81

5.1 WLS Regression Results of Treatment feature for projects ZUEP and MIPL 104
LIST OF FIGURES

1.1 Hidden variables 5

2.1 Boxplot of annual average NDVI values observed at World Bank project locations over time and worldwide 16

2.2 Boxplot of slope of a linear regression model fitted with annual average NDVI data 17

2.3 Change of annual average NDVI values 17

2.4 Geographical locations 19

2.5 World Bank projects (2010-2012) 19

2.6 Distance to commercial rivers and distance to all rivers 21

2.7 Locations of outliers in distance to commercial rivers and distance to rivers 22

2.8 Distance to Roads 23

2.9 Locations of outliers in boxplot of distance to roads 23

2.10 Accessibility to urban areas 25

2.11 Locations of outliers in boxplot of travel time 26

2.12 Population Density 27

2.13 Outliers locations 28

2.14 Slope and elevation 30
2.15 Slope and elevation outliers
2.16 Intercept and slope values of fitted linear regression models for time series data on annual average, max, and min temperatures at project locations
2.17 Geographical locations of projects with temperature time series data whose regression models have intercept or slope values shown as outliers in their boxplots
2.18 Histograms for intercept and slope values without outliers for regression models of temperature time series data
2.19 Intercept and slope values of fitted linear regression models for time series data on annual average, max, and min precipitation at project locations
2.20 Geographical locations of projects with precipitation time series data whose regression models have outliers for intercept or slope values
2.21 Histograms for intercept and slope values without outliers for regression models of precipitation time series data
2.22 Starting years of World Bank projects
2.23 Boxplot of funding amounts for projects and the geographical locations of outliers
2.24 Correlation among all covariates
2.25 Covariates overlap between treatment and control
3.1 Estimated treatment effects for randomized assignment, $e(x) = 0.5$
3.2 MSE changes with data size
3.3 CF calculated distributions of treatment effect estimates for specific projects: (a) Saint Lucia Hurricane Tomas Emergency Recovery Loan; (b) Sustainable Tourism Development Project; (c) Emergency Infrastructure Reconstruction Project.

3.4 (a) Causal effect distribution of all World Bank projects combined and (b) separated by continents

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

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

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

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

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

4.6 MSE changes with data size for CT and CF, left figure $k=1$ and right figure $k=8$.

5.1 Distance in a causal tree

5.2 Histogram of values contributed by trees in the CF for ZUEP to compute $\hat{\tau}(x)$
5.3 Geographic locations of a sampled subset of projects used in the estimation of \( \hat{\tau}(x) \) for ZUEP

5.4 ZUEP feature frequency in CF

5.5 Zhejiang Urban Environment Project: Weights derived from distance values after scaling with a kernel function

5.6 Histogram of values contributed by trees in the CF for MIPL to compute \( \hat{\tau}(x) \)

5.7 Neighbors used in the weighted linear regression for MIPL

5.8 MIPL feature frequency in CF

5.9 Distances scaled by kernel function for MIPL

5.10 Distribution of change of nights lights

5.11 Causal effect of World Bank projects to change of night lights
Chapter 1

Introduction

1.1 Correlation and Causality

Both correlation and causality can describe a relation between two variables, but causality is different from correlation in that it has direction.

Correlation measures how strong two variables are linearly related. When we know one of them, we can also predict another if they have a strong correlation. We often measure this relation using the correlation coefficient. Their relationship can be causal, but it can also be that they simply happen together by coincidence. For instance, people who buy a mouse always buy a mousepad at the same time. Therefore, a correlation exists between buying a computer mouse and buying a mousepad. In this case, buying a mouse and buying a mousepad have a correlation, but we cannot draw the conclusion that buying a mousepad is caused by buying a computer mouse.

Causality describes an even more specific relation because it has direction between cause and effect. For instance, “His headache went away because he took an aspirin,” or “He got lung cancer because he smoked a lot.” In the above sentences, the word “because” indicates the causal relation between the headache and aspirin, lung cancer, and smoking. But the
word “because” is not necessarily used in a formally correct way in daily language or even articles. As an example, in the sentence “His feet are large because he is tall,” we know that height is not the cause of large feet, so the word “because” does not correctly describe the relation.

It is very important to find the causality relation as it can help us to better understand problems. As shown from the above example, people may not express the causal relation properly, and we always use causality with our own experience. However, if we can find a true causal relation, we can benefit a lot from it. If we can change the cause, then we can change the result. For example, if we know the cause of a disease, we have an opportunity to prevent it. But for correlation, when we change one variable, we may or may not change the other variable. As such, the difference is that causality has direction.

In [41], Imbens et al. view causality as an action that is applied to a unit. Under their definition, an action can be a treatment, intervention, or manipulation, and a unit can be a person, a firm, or a project, in our case, at a particular point in time. The same person at a different time is a different unit under this definition. For example, we want to investigate the causality between aspirin and headache. A person took an aspirin when he got a headache, and we recorded if his headache was reduced at some later time. He got a headache again, and he did not take an aspirin at this time. Therefore, we cannot draw the conclusion of the causality between aspirin and the headache, because the person at these different times is not the same unit.

This thesis considers only two actions: treated and control. In our scenario, treated action means investing more money into the project, while control actions mean putting a small amount of money into the project.
1.2 Causal Inference in Observational Studies

Causal inference is an active and important research area across many disciplines. In economics, epidemiology, medicine and political science, matching is often used to estimate causal effect, as in [73]. In [67], they demonstrated how to find the directions of two related variables. In [70], they aimed at solving the confounder problem. In [56], Judea proposed a solution based on the Bayesian network. In [61], [41] and [62], Imben and Rubin proposed a potential outcome framework to estimate the causal effect given the treatment.

In this thesis, we are interested in answering the research question: What is the impact of a World Bank project on the change of environment? In chapter 2, we detail the World Bank dataset and how it is collected by AidData. Also, we describe why and how we transform the dataset such that we can address the research question within the potential outcome framework and with tree-based causal inference techniques.

1.2.1 Potential Outcome Framework (Rubin-Neyman Causal Model)

In the potential outcome framework as described in [62], each unit has two potential outcomes given a binary treatment. \( Y(1) \) is the outcome if the unit had been treated, while \( Y(0) \) is the outcome had the unit not been treated. For example, Bob is a person who has a headache and we consider a treatment with aspirin. He can either take an aspirin or not. There are two potential outcomes: \( Y(1) \) had he taken an aspirin, and \( Y(0) \) had he not taken an aspirin. For any given individual, we can only observe one of the two potential outcomes; the unobserved outcome is the counterfactual. If we have both of these two potential outcomes, we can estimate the causal effect of aspirin as \( Y(1) - Y(0) \). However, we cannot observe both factual and counterfactual at the same time. Therefore, a causal inference problem becomes a missing data problem, as we can only observe one potential outcome, either \( Y(1) \) or \( Y(0) \).
In the potential outcome framework, given \( Y_i(1) \) and \( Y_i(0) \), we can estimate the causal effect for each individual unit as \( ITE_i = Y_i(1) - Y_i(0) \) for any unit \( i \). We can also estimate the average treatment effect as \( ATE = E[Y_i(1) - Y_i(0)] \) for some population or set of units.

Methods to estimate a causal effect depend on a number of assumptions, including ignorability, overlap, and the absence of spillover effects. We discuss these assumptions in the next section.

1.2.2 Common Assumptions in Causal Inference

In an experimental setting, one can randomly assign a treatment to generate a treated and control group with similar units and then apply causal inference techniques. However, for observational studies, it is a common case that a unit was or was not treated for some reason and the treatment assignment can not be freely chosen. So, to estimate causal inference for observational data, we have to make several important assumptions, as described in [59], including ignorability and common support. An additional assumption, the Stable Unit Treatment Value Assumption (SUTVA), is described in [60].

1.2.2.1 Ignorability or Absence of Unobserved Confounders

\[ W \perp Y \mid \mathcal{X} \quad (1.1) \]

In equation 1.1, treatment assignment \( W \) is independent of the potential outcomes, the \( Y \), given a condition on covariates \( \mathcal{X} \). The ignorability assumption means there are no unobserved confounders. A confounder is a covariate that can influence both outcome and treatment assignment. For the covariates in this condition, they are not affected by the treatment. This condition implies that we observe all the confounders.

For example, as shown in figure 1.1, observed covariates include age, gender, and job among others excluding genetic variables. Genetic variables are unobserved. The treatment
is smoking, and the outcome (or response) is lung cancer. We want to investigate the causal relation between smoking and lung cancer. We cannot draw the conclusion that smoking causes lung cancer, because lung cancer can be caused by genetic variables. Because there is a correlation between smoking and genetic variables, we cannot tell which one causes the lung cancer.

Figure 1.1: Hidden variables

1.2.2.2 Common Support or Overlap

\[ 0 < P(W = 1 \mid X = x) < 1 \] (1.2)

The common support condition, which is shown in equation 1.2, is an assumption that measures overlap of covariates distributions between treated and control units. The condition excludes values of 0 and 1 for \( P(W = 1 \mid X = x) \). It implies that there are similar units in both the treated and control group. We need to check the overlap assumption for any dataset before we estimate the causal effect within the potential outcome framework. For example, if we assess the balance of the matched treated and control data, then we have to compare the joint distribution of all covariates for the matched treated and control group.
which creates a challenge for high dimensional data. Therefore, in practice, the propensity score is used as a balancing score to access the balance of the treated and control groups.

1.2.2.3 Stable Unit Treatment Value Assumption (SUTVA)

The Stable Unit Treatment Value Assumption (SUTVA) assumes there is no spillover problem when estimating the causal effect, which means the outcome or response of a unit can only be affected by its own treatment, not impact by treatment of other units. More formally, we define SUTVA as follows.

**Definition 1** The Stable Unit Treatment Value Assumption (SUTVA) holds if for any unit \( i \), its potential outcome \( Y_i \) can only be impacted by its treatment \( W_i \), and \( Y_i \) will not be affected by \( W_j \), where \( j \neq i \). \( Y_i \) is only impacted by the covariates \( X \).

For World Bank projects, this is a strong assumption for the research problem to estimate the causal effect of a given World Bank project on the environment. Obviously, the World Bank project may be affected by another project. But for the sake of simplicity, this thesis holds the SUTVA assumption.

1.3 Heterogeneity in Causal Effects

For the World Bank dataset, we see a large variety in the kinds of projects funded by the World Bank. So, besides the average treatment effect that measures the effect of all the projects on the environment, we are more interested in quantifying the heterogeneous causal effect, as we know that different kinds of projects produce different causal effects on the environment. This will enable us to calculate the causal effect for each project in a refined manner.
In the literature, particularly in [22], many methods have been proposed to estimate the heterogeneous causal effect. Given a propensity score for each unit, we can use a different matching method to estimate its causal effect. There are many matching algorithms, such as nearest neighbor matching, kernel, and linear matching. Although these methods work well with a small number of covariates, they do not work well with a larger set of covariates.

Moreover, there are tree-based and forest-based methods in the literature as well, such as in [35], [82], and [5]. In [5], Athey et al. proposed the causal tree and causal forest methods to estimate heterogeneous causal effects. In this thesis, we use the causal tree and causal forest proposed by Athey et al.; however, we depart from simulation studies to use a real World Bank dataset. We refine the causal tree and causal forest approaches such that we can apply them to our data.

1.3.1 Decision Tree

Decision Tree [57] is a supervised machine learning technique for classification and regression. A decision tree is a binary tree in which each node can be split into two children, and each node contains different data. Given a split criterion, a tree will grow until the node cannot split anymore. An optimal split will grow the tree until no more benefit is to be had from continued splitting. Gini information gain and mean square error reduction are often used as the splitting criteria in building a regression tree.

Besides the splitting criterion, there are various parameters that control the growth of a tree, including the leaf size. If a deep tree is needed, then the leaf size can be set to a small number; if a shallow tree is considered preferable, one can set the leaf size to be a large number.

After a tree is constructed with the training data, we can predict the value or class of the test data. We can start with the root node and choose the path to a leaf node according
to the feature and threshold along the path, and we can use the leaf node value or class as the result for test data.

### 1.3.2 Causal Tree

A causal tree shares the same structure of a decision tree: it is a binary tree. However, causal trees are different from decision trees in three aspects:

1) **Splitting criterion.** Although the splitting criterion of a causal tree uses a similar idea as the mean square error (MSE) splitting criterion often used in a decision tree, the ground truth is missing in a causal tree. We describe the details in chapter 3.

2) **Treated and control data in each node.** As we have both treated and control units, we must calculate the average treatment effect (ATE) for each node. Data in the decision tree do not distinguish treated and control data. For the World Bank data set, as we want to estimate the causal effect for each project, we need to make sure there are both treated and control units in each node, and we have to avoid “extreme” nodes. We call a node with either only treated units or only control units an “extreme” node. We investigate such splitting criteria with synthetic data, and we find that extreme nodes are produced in a regular manner and for the treated units they contain, no causal effect can be estimated. To prevent the generation of extreme nodes, we develop new splitting criteria as shown in chapter 3. This extends the work of [5], where extreme nodes are not prevented but lead to cases where no causal effects are estimated.

3) **Results.** For decision trees, we build a tree with training data and we do regression or classification for the test data. However, for the causal tree, we do not have training data. A causal tree can partition feature spaces into smaller subspaces, and the causal effect within that space is constant. We can estimate the causal effect for each project, which can be either treated or control.
1.4 Other Causal Inference Approaches

While we adopt Rubin’s approach as it enables the causal trees explored here, there are other approaches such as Granger causality and Judea Pearl’s causal model.

1.4.1 Judea Pearl’s Causal Model

Judea Pearl’s causal model [55] represents covariates as nodes in a directed acyclic graph (DAG), and the edges are derived from the structural equations. The intervention is represented by the do(x) operator, and the back door criterion is used to select covariates which are used to calculate the causal effect. To estimate the causal effect of X on Y, where X and Y are nodes in the DAG (which is also called the causal graph), the back door criterion can find the set Z, which d-separates the path between X and Y. If there are no confounders, then X \perp Y \mid Z. The idea behind the back door criterion is that for the path from X to Y, we can observe causality between X and Y. However, the back door path will carry spurious association between X and Y. The back door criterion is used to find the set of nodes that block all the paths from X to Y, that is, from treatment to response. To represent intervention, do(x) is proposed. For example, to estimate the causal effect of X on Y, P(y \mid do(x)) is the causal effect. Pearl proved that P(y \mid do(x)) = \sum_z P(z)P(y \mid x, z). do(x) means setting X to a number x in the structured equations and removing all the edges which include X in the causal graph. Comparing the causal effect with the potential outcome framework,

$$\tau = E[Y_i \mid do(x = 1)] - E[Y_i \mid do(x = 0)]$$ (1.3)

In Judea Pearl’s model, the structural equations are non-linear, nonparametric equations. For example, x = f(y, \theta), where x, y are covariates, \theta is noise. We do not need to
exactly know the form of the function. y is on the right side, and in the causal graph, there is an edge between x and y with an arrow emitting from y and pointing to x.

1.4.2 Granger Causality

Granger causality [34] can be used to find if two time series covariates have a causal relation. It can be tested by a NULL hypothesis test that x does not cause y, both x and y are time series with a stationary process. Details can be found in [88]. In Granger causality, causality is viewed in the prediction aspect. Knowing the history of x will not help to improve the prediction of y. So, then, we say that x is not granger causing y. Formally, we have the following:

\[ P(Y(t + 1) \mid I(t)) \neq P(Y(t + 1) \mid I_{-x}(t)) \]  

(1.4)

where I(t) is all the information in history and I_{-x}(t) is all the history information, excluding the history of x. If the above equation holds, then we say x causes y. In the null hypothesis, we can test with different lags to see which lag is significant. Besides the Granger causality, in [19], they used the difference-in-differences approach to estimate the causal effect with time series data. They used a diffusion-regression state-space model to estimate the control group. For example, they can estimate the effects of government or marketing policy, such as an ads campaign for a company such as Google. This method can only estimate the group causal effect. We are interested in heterogeneous effects and this method can only give the overall causal effect.
1.5 Interpretable Model

We can achieve very good accuracy using very complex machine learning models such as random forest and CNN. In this thesis, the causal forest enables us to accurately estimate the causal effect for any World Bank project. However, the causal forest results are also hard for humans to interpret as well. We have noticed a growing literature on interpreting complex machine learning models. In [26] and [2],[4], they discuss the importance of the interpretable machine model to AI safety. In [58], the LIME approach fits the data to be explained and its neighbors to a linear regression model and interpret the result with the fitted model. In [42] and [2], they use an influence function for the explanation. In [45], the author gives a comprehensive discussion of interpretability in machine learning research.

1.6 Overview

The rest of this dissertation is structured as follows. Chapter 3 discusses the heterogeneous causal effect of World Bank projects. This chapter is taken almost verbatim from a paper published in [90]. Chapter 4 shows results of a simulation study to analyze the performance of causal trees and causal forest methods with a stochastic model that uses distributions fitted to covariates in the World Bank data. It is taken almost verbatim from a paper we published in [91]. Chapter 5 discusses the explanation model of the causal forest, which is in the form of a paper submitted for publication. We conclude with a summary of our findings in Chapter 6.
In this chapter, we detail the data used in the thesis. The World Bank dataset was collected from AidData, and table 2.1 lists the data and its individual source. The focus of this thesis is to answer the research question: How much does a World Bank project influence the environment? In order to address this question with data analysis, the data needs to represent measurable characteristics of the environment, of World Bank projects, as well as of other potentially confounding factors that may influence the environment. In [86], they demonstrated that every aspect of our lives depends upon plants and trees. Our health, economy, and environment are all affected when drastic changes occur to vegetation. Therefore, we represent the environment with vegetation and we use the Normalized Difference Vegetation Index (NDVI) as the metric to measure the density of vegetation. In the dataset, there are covariates accounting for civilization factors, including distance to commercial rivers, distance to roads, population density, and travel time to large cities. This is based on the understanding that human economic activity influences the environment a lot and our data should reflect factors that relate to farming, urbanization, or industrialization. Also, there are covariates representing geological and general climate information, including distance to rivers, elevation, slope, temperature, and precipitation. The covariates relate to the
availability of water and sunlight which obviously influence vegetation. Our time scale is in years, so the idea is to distinguish between wet and dry years with above or below average precipitation or hot and cold years respectively. For the World Bank projects themselves, we have key project characteristics such as the time line and funding amounts as well as their geographical location. As funding triggers human activity on the ground, we interpret the funding amounts as an indication of how much of a difference a World Bank project will make at its location. The precise geographical location is also crucial to combine data from different sources in a meaningful way.

Given the covariates listed in table 2.1, we performed a basic characteristics analysis for each of the covariates. These covariates can potentially affect vegetation such as temperature, precipitation, and distance to rivers. Among them, some are time series data. For example, precipitation is a time series with data collected from the years 1982 to 2014. Some covariates do not have a time component, such as geolocation: the project longitude and latitude will not change over time.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest Cover</td>
<td>NASA Long Term Data Record measurements of vegetative cover</td>
</tr>
<tr>
<td></td>
<td><a href="http://ltdr.nasa.com/cgi-bin/ltdr/ltdrPage.cgi">http://ltdr.nasa.com/cgi-bin/ltdr/ltdrPage.cgi</a></td>
</tr>
<tr>
<td>World Bank Project Locations</td>
<td>Double-blind geocoded information on the geographic location of each World Bank project</td>
</tr>
<tr>
<td></td>
<td><a href="http://aiddata.org/level1/geocoded/worldbank">http://aiddata.org/level1/geocoded/worldbank</a></td>
</tr>
<tr>
<td>Distance to Rivers and</td>
<td>The calculated average distance to rivers</td>
</tr>
<tr>
<td>Distance to Roads</td>
<td>Distance to nearest road</td>
</tr>
<tr>
<td>Elevation, Slope</td>
<td>Elevation and slope data measured from the Shuttle Radar Topography Mission</td>
</tr>
<tr>
<td></td>
<td><a href="http://www2.jpl.nasa.gov/srtm/">http://www2.jpl.nasa.gov/srtm/</a></td>
</tr>
<tr>
<td>Accessibility to Urban Areas</td>
<td>European Commission Joint Research Centre estimation of urban travel times.</td>
</tr>
<tr>
<td>Population Density</td>
<td>Center for International Earth Science estimation of population density, derived from Nighttime Lights</td>
</tr>
<tr>
<td>Air Temperature, Precipitation</td>
<td>University of Delaware Long term, global temperature and precipitation data interpolated from weather station measurements.</td>
</tr>
<tr>
<td></td>
<td><a href="http://climate.geog.udel.edu/climate/">http://climate.geog.udel.edu/climate/</a></td>
</tr>
</tbody>
</table>

**Table 2.1:** Covariates of the data sets for World Bank projects
2.1 Vegetation

Satellite data are often used to measure vegetation, which has the advantage of the longest continuous time record. Therefore, we decided to take advantage of satellites to measure the vegetation globally, as World Bank projects spread around the world. NDVI measures the relative absorption and reflectance of red and near-infrared light from plants to quantify vegetation on a scale of -1 to 1, with vegetated areas falling between 0.2 and 1, as described in [29]. While the NDVI does have a number of challenges, including a propensity to saturate over densely vegetated regions, the potential for atmospheric noise (including clouds) to incorrectly offset values, and the reflectance from bright soils providing misleading estimates, the popularity of this measurement has led to a number of improvements over time to offset many of these errors. This is especially true of measurements from longer-term satellite records, such as those used in this analysis, produced from the MODIS and AVHRR satellite platforms [54].

The dataset provides annual average values for NDVI such that seasonal effects are averaged out. Figure 2.1 shows boxplots of average annual NDVI values of all project locations for each year from 1982 to 2014. The remote sensing images from satellites ranged from 250m to 4km. The details are shown in table 2.1. The mean values are non-negative for all projects over all years, and typical values are around 0.2, which is a lower bound for areas with vegetation. In figure 2.1, NDVI values in the year 2000 have outliers of a much larger value than in the other years, so these projects may have been located in areas with a lot of vegetation. Most of the projects are located in places with an NDVI value of less than 0.5.
While causal inference techniques can be applied to time series data where each year is a covariate on its own, it is then difficult to separate pre-treatment and post-treatment data for projects that started in different years. We decided to aggregate the time series data into two covariates, namely the intercept and slope of a fitted linear regression model, for the years before a project started at a particular location. We use the intercept and slope as the covariates to represent NDVI values that are not affected by a project. In this way, our data includes another covariate that may help explain environmental changes much like precipitation and temperature. There are more refined time series models than a linear regression such as the Auto-Regressive Time Series Model and the Moving Average Time Series Model shown in [36], but we leave those for future work.

Figure 2.2 shows the distributions of slopes of a linear regression model fitted by time series data of NDVI values that starts in 1982 and ends with the year before each project started. The projects started in various years from 2000 to 2011. Approximately 75% of all projects have an upward trend in NDVI values across this time period.
Figure 2.2: Boxplot of slope of a linear regression model fitted with annual average NDVI data

(a) $\Delta NDVI$: Change of annual average NDVI values before and after a project started

(b) $\Delta NDVI$ vs. funding

Figure 2.3: Change of annual average NDVI values

In the thesis, $\Delta NDVI$ is calculated by the difference between the average of the annual
NDVI values observed at the project location before and after the projects started. So, we split the time series of NDVI values into two, one for the years before the project started and one for the years after that and consider the difference between the average values of the two series. We use $\Delta NDVI$ to describe the change in vegetation, the change in environment, that is observed in the time when a World Bank project starts and at the location of that project. Figure 2.3 shows the distribution of $\Delta NDVI$ values for all projects as a box plot and also in relation to the funding amounts of the projects. The former shows that $\Delta NDVI$ observations are small in value with an average close to zero. The latter shows that the variability of $\Delta NDVI$ decreases with the amount of funding but so does the total amount of projects per funding level. For the causal inference analysis, the question is which part of the observed change in NDVI is attributable to the presence of a World Bank project.

### 2.2 Geographical Locations

The second primary dataset used in this analysis measures where World Bank projects were geographically located. Geographical locations are crucial to combine information from different data sources, for example to extract NDVI values from satellite data for the location of a particular World Bank project. Geographical location data is used to calculate the distance from a project location to some other entity of interest, for instance the distance to the nearest urban center or the distance to a body of water. In our analysis, longitude and latitude are covariates that allow for a spatial clustering of projects. However, spatial proximity is not necessarily a good metric for measuring similarity between projects. For example, if two projects are in close proximity but located in two different countries, they may be quite different because some civilization factors such as population and accessibility to urban areas can be completely different.
In figure 2.4, we observe that World Bank projects are widely distributed around the world. Asia alone has half of all World Bank projects. Figure 2.4 demonstrates that the majority of the projects are located above the equator. For longitude, most projects can be partitioned into three parts: Asia, Africa and South America (Figure 2.5).

This geolocation dataset was produced by [3], relying on a double-blind coding system where two experts independently assign latitude and longitude coordinates, precision
codes, and standardized place names to each geographic feature. Disagreements are then arbitrated by a third party.

2.3 Distance to Rivers

In [43], they demonstrate that over 50% of the world’s population live within 3km of fresh water. Rivers play important roles in agriculture irrigation, navigation, and hydropower generation. However, human development brought pollution and degradation as described in [33]. We especially consider large rivers that can serve as part of an overall transportation infrastructure (commercial rivers) for their economic importance. Rivers of any size are taken into account for their importance to civilization but also as a characteristic of the overall geographic and environmental setting.

In figure 2.6, we see the distributions of both distance to commercial rivers and distance to any rivers. From the boxplot, we observe that there are outliers with very large values. The locations of these outliers is shown in figure 2.7. We observe that some projects located in islands are far from commercial rivers, and some projects in Africa are far from rivers. Figure 2.6 (b) shows the average distance from a location to commercial rivers after removing outliers, and the distance to commercial rivers ranges from 0 to 400km. (c) shows the average distance from a location to any river after removing outliers. The projects are all close to rivers, at distance ranging from 0 to 6km. In table 2.2, we observed that the max distance is 16000km. The extreme large values may be caused by the procedure used to calculate the distance by AidData or caused by a lack of data for rivers. Given the outliers, there are several options. We can replace the outliers with N/A, we can delete the covariate or we could choose to remove the projects in the dataset. Because the river database may not be complete, we can also update the database and recalculate the distance. Moreover, we can adjust the values, for example, if a value is larger than 400km, we can use 400km
to replace it with the understanding that 400km is a distance to a commercial river that is too far to matter for the project. Finally, we can also analyze their potential impact on the outcome of the causal inference analysis.

![Boxplot of distance to commercial rivers and distance to rivers](image1)

(a) Boxplot of distance to commercial rivers and distance to rivers

![Histogram of distance to commercial rivers after removing outliers](image2)

(b) Histogram of distance to commercial rivers after removing outliers

![Histogram of distance to rivers after removing outliers](image3)

(c) Histogram of distance to rivers after removing outliers

**Figure 2.6:** Distance to commercial rivers and distance to all rivers

The causal tree and causal forest are not sensitive to the outliers since covariates are only used to find optimal features and thresholds for splitting a node in growing a tree. The features are used for separation rather than being used to calculate the causal effects.
However, for a linear regression model, outliers may impact the final result.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.93e+02</td>
<td>7.00e+02</td>
<td>0.00e+00</td>
<td>3.041e+02</td>
<td>7.80e+02</td>
<td>1.68e+02</td>
<td>1.60e+04</td>
</tr>
</tbody>
</table>

Table 2.2: Distance to large rivers with a potential commercial use (in km)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.51e+01</td>
<td>6.63e+02</td>
<td>0.00e+00</td>
<td>1.00e+00</td>
<td>1.41e+00</td>
<td>3.00e+00</td>
<td>1.60e+04</td>
</tr>
</tbody>
</table>

Table 2.3: Distance to rivers (in km)

![Map of commercial rivers](image1.png)
![Map of rivers](image2.png)

(a) Locations of outliers in distance to commercial rivers
(b) Locations of outliers in distance to rivers

Figure 2.7: Locations of outliers in distance to commercial rivers and distance to rivers

2.4 Distance to Roads

As shown in [9], roads contribute to development. Roads bring economic and social benefits, but they also come with pollution and deforestation.
(a) Boxplot of distance to roads  
(b) Histogram of distance to roads after removing outliers

Figure 2.8: Distance to Roads

Figure 2.9: Locations of outliers in boxplot of distance to roads
<table>
<thead>
<tr>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.01e+01</td>
<td>3.58e+02</td>
<td>0.00e+00</td>
<td>0.00e+00</td>
<td>1.00e+00</td>
<td>2.00e+00</td>
<td>7.55e+03</td>
</tr>
</tbody>
</table>

**Table 2.4**: Distance to roads (in km)

As shown in figure 2.8 and table 2.4, most projects are close to roads and the median distance is 1km. This is not surprising as the World Bank aims to overcome poverty. Most of the projects should therefore be close to people and people live near roads. There are some projects that are seen as outliers in figure 2.9 based on the distance to roads. The outliers have distances over 5000 km in the boxplot. As we discussed for distance to rivers, the outliers can be produced by the roads database or the procedure used to calculate the nearest roads. Although the causal tree and causal forest method are not sensitive to the outliers, it is important to consider them for a linear regression model. We can adjust the outliers to a fixed value (for example 1000km), update the road database or apply a better model to calculate the distance to roads. We show the geolocations of the outliers in figure 2.9. We observed that these projects are spread over Asia, Africa and America. Most of these projects are transportation projects, such as the Second Eastern Indonesia Region Transport Project, which is important for an area far away from roads with poor transportation.

Moreover, if a project is far away from roads, it may imply that the vegetation of that area is mostly impacted by the environment, such as temperature or precipitation, and less impacted by humans. Therefore, distance to roads is an important covariate that can impact the environment in the area as transportation projects may have negative impacts on the NDVI value.
2.5 Accessibility to Urban Areas

Accessibility to urban areas is to control for the potential of human-dominated forest loss in contrast to other factors. Figure 2.10 and table 2.5 show the travel time from the project locations to nearby big cities. The travel time data is based on the source shown in 2.1. In the travel time database, each pixel in the map is mapped to a cost time and we can calculate the travel time given any project location. The median travel time is about 1 hour, the outliers are values higher than 350 minutes. In figure 2.11, the projects are mostly located in Asia and Africa. For example, projects located in islands of Asia are far from cities.

Figure 2.10: Accessibility to urban areas
Figure 2.11: Locations of outliers in boxplot of travel time

<table>
<thead>
<tr>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.30e+02</td>
<td>2.52e+02</td>
<td>0.00e+00</td>
<td>1.80e+01</td>
<td>6.60e+01</td>
<td>1.55e+02</td>
<td>4.65e+03</td>
</tr>
</tbody>
</table>

Table 2.5: Travel time (in minutes)

2.6 Population Density

In [27], the authors demonstrate that population is closely related to environment. Human land use can be an important factor on the change of environment. For example, for a lower range of population density values, one can expect to see the impact of different types of farming and land use. For a higher range of values, one can expect to see a
change in vegetation due to urbanization with vegetation free spaces for buildings and roads. As shown in figure 2.12, projects are located in places with various population densities. Population density of the years 1990, 1995 and 2000 do not change a lot. In table 2.6, the median of population density is about 200 persons per square kilometer. However, some projects are located in places with large populations and high population densities. As shown in figure 2.13, most outlier projects are located in China and India.

(a) Boxplot of population in year 1990, 1995, 2000
(b) Histogram of population in year 1990 after removing outliers

(c) Histogram of population in year 1995 after removing outliers
(d) Histogram of population in year 2000 after removing outliers

**Figure 2.12: Population Density**
Table 2.6: Population density in different years (in persons per square km)

<table>
<thead>
<tr>
<th>year</th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>7.62e+02</td>
<td>2.40e+03</td>
<td>0.00e+00</td>
<td>5.20e+01</td>
<td>1.55e+02</td>
<td>5.40e+02</td>
<td>7.89e+04</td>
</tr>
<tr>
<td>1995</td>
<td>8.48e+02</td>
<td>2.70e+03</td>
<td>0.00e+00</td>
<td>5.52e+01</td>
<td>1.72e+02</td>
<td>5.86e+02</td>
<td>8.59e+04</td>
</tr>
<tr>
<td>2000</td>
<td>9.29e+02</td>
<td>2.99e+03</td>
<td>0.00e+00</td>
<td>5.98e+01</td>
<td>1.86e+02</td>
<td>6.29e+02</td>
<td>9.28e+04</td>
</tr>
</tbody>
</table>

Figure 2.13: Outliers locations

(a) locations of outliers of population density in 1990
(b) locations of outliers of population density in 1995
(c) locations of outliers of population density in 2000

Population density calculation is based on night light satellite data, and therefore errors...
exist between the real number and the estimated value. According to the World Bank survey, the largest population density is Macau in China for 2016 (20,204 persons per square kilometer). There are several projects with a population density larger than 20,000 in the World Bank dataset. We can transform the outliers to value 20,000 instead given the population density in Macau.

### 2.7 Slope and Elevation

In [52], they modeled the relationship between elevation and vegetation. Topography, which include both slope and elevation, is an important factor in vegetation growth. Elevation and slope are both key factors in controlling the vegetation in mountain areas.

For example, areas with high slopes can result in decreased deforestation due to events other than World Bank projects interventions - i.e., flash floods, which is why it’s important to be included in the dataset.

<table>
<thead>
<tr>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.49e+00</td>
<td>4.06e+00</td>
<td>0.00e+00</td>
<td>0.28e+00</td>
<td>0.88e+00</td>
<td>2.73e+00</td>
<td>4.31e+01</td>
</tr>
</tbody>
</table>

**Table 2.7**: Slope (in degree)

<table>
<thead>
<tr>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.00e+02</td>
<td>8.00e+02</td>
<td>-4.20e+01</td>
<td>4.70e+01</td>
<td>2.61e+02</td>
<td>8.61e+02</td>
<td>5.62e+03</td>
</tr>
</tbody>
</table>

**Table 2.8**: Elevation (in m)

Figure 2.14 (a) and (b) show the boxplot of both slope and elevation. Figure (c) and (d) show the histogram of slope and elevation after removing outliers. We observe that
most projects have a small slope between 0 to 6 degrees and an elevation ranging from 0 to 2km. In figure 2.15, it is easy to find that the outliers are projects located in area with mountains. To deal with the outliers in slope and elevation, their values are accurate and we can leave them the same.

Figure 2.14: Slope and elevation
2.8 Temperature

Temperature is an important factor to be considered because it can mediate tree growth in different biomes as described in [85]. Temperature values are available as a time series of annual average, annual minimum, and annual maximum temperatures for each project location. Temperatures naturally fluctuate over time. By considering annual values, we already removed the effect of seasons from the temperature data. The data is able to differentiate hot from cold years as well as recognizing years with an extreme temperature condition. For each individual time series, we fit a linear regression model and extract its intercept and slope. In this way, we have three intercept values for each project: one for average temperature, one for maximum temperature, and one for minimum temperature. In the same way, we obtain values for the slope of minimum, maximum, and average temperatures. In figure 2.16 and tables 2.9 and 2.10, we show some basic statistics for the different intercept and slope values we derive for the projects.
Figure 2.16: Intercept and slope values of fitted linear regression models for time series data on annual average, max, and min temperatures at project locations

Table 2.9: Statistics for the intercept of a fitted linear regression for average, min, and max temperature time series data

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg</td>
<td>1.75e+01</td>
<td>5.63e+00</td>
<td>-6.59e+00</td>
<td>1.35e+01</td>
<td>1.90e+01</td>
<td>2.16e+01</td>
<td>3.28e+01</td>
</tr>
<tr>
<td>max</td>
<td>2.14e+01</td>
<td>6.18e+00</td>
<td>-2.72e+01</td>
<td>2.00e+01</td>
<td>2.24e+01</td>
<td>2.46e+01</td>
<td>3.35e+01</td>
</tr>
<tr>
<td>min</td>
<td>9.99e+00</td>
<td>1.07e+01</td>
<td>-6.25e+01</td>
<td>6.73e-01</td>
<td>1.13e+01</td>
<td>1.81e+01</td>
<td>6.51e+01</td>
</tr>
</tbody>
</table>

Table 2.10: Statistics for the slope of a fitted linear regression for average, min, and max temperature time series data

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg</td>
<td>2.84e-02</td>
<td>1.98e-02</td>
<td>-7.32e-02</td>
<td>1.47e-02</td>
<td>2.69e-02</td>
<td>4.08e-02</td>
<td>1.41e-01</td>
</tr>
<tr>
<td>max</td>
<td>2.69e-02</td>
<td>2.73e-02</td>
<td>-9.42e-02</td>
<td>7.37e-03</td>
<td>2.21e-02</td>
<td>4.39e-02</td>
<td>1.97e-01</td>
</tr>
<tr>
<td>min</td>
<td>4.75e-02</td>
<td>2.23e-01</td>
<td>-2.04e+00</td>
<td>1.71e-03</td>
<td>2.09e-02</td>
<td>4.50e-02</td>
<td>5.77e+00</td>
</tr>
</tbody>
</table>
Figure 2.17: Geographical locations of projects with temperature time series data whose regression models have intercept or slope values shown as outliers in their boxplots
Figure 2.18: Histograms for intercept and slope values without outliers for regression models of temperature time series data.
We find that most of the projects are located in areas with positive temperature trends according to the slopes. For the intercepts, we can observe that average temperature, maximum temperature and minimum temperature have medians around 19, 22 and 11 degrees, respectively. In figure 2.18, we show the histogram for the slope and intercept. We observe that heterogeneity exists in both of them and they have different impacts on the environment. In figure 2.17, we find most of the outliers (for both slopes and intercepts) around the equator, which leads to a big value for intercepts.

2.9 Precipitation

Precipitation obviously impacts vegetation. Just as for temperature, precipitation values are available as a time series of annual average, annual minimum, and annual maximum precipitation for each project location. Precipitation naturally fluctuate over seasons. By considering annual values, we already removed the effect of seasons from the precipitation data. The data is able to differentiate dry from wet years as well as recognizing years with an extreme precipitation condition. For each individual time series, we fit a linear regression model and extract its intercept and slope. In this way, we have three intercept values for each project: one for average precipitation, one for maximum precipitation, and one for minimum precipitation. In the same way, we obtain values for the slope of minimum, maximum, and average precipitation. In figure 2.19, we have both slope and intercept distribution for average, max, and min precipitation. Tables 2.11 and 2.12 show some basic statistics for these distributions.

In figure 2.21, we observed that about half of the projects have a positive trend, and the other half have a negative trend. For the intercept, projects have various values, which indicate a lot of heterogeneity in precipitation.
Figure 2.19: Intercept and slope values of fitted linear regression models for time series data on annual average, max, and min precipitation at project locations

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg</td>
<td>9.46e+01</td>
<td>5.58e+01</td>
<td>0.00e+00</td>
<td>5.30e+01</td>
<td>8.58e+01</td>
<td>1.19e+02</td>
<td>5.43e+01</td>
</tr>
<tr>
<td>max</td>
<td>1.72e+02</td>
<td>1.04e+02</td>
<td>0.00e+00</td>
<td>9.18e+01</td>
<td>1.65e+02</td>
<td>2.30e+02</td>
<td>9.09e+01</td>
</tr>
<tr>
<td>min</td>
<td>2.24e+01</td>
<td>1.76e+01</td>
<td>-4.06e+00</td>
<td>1.11e+01</td>
<td>1.66e+01</td>
<td>2.57e+01</td>
<td>1.44e+01</td>
</tr>
</tbody>
</table>

Table 2.11: Statistics for the intercept of a fitted linear regression for average, min, and max precipitation time series data

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg</td>
<td>5.00e-02</td>
<td>9.61e-01</td>
<td>-7.12e+00</td>
<td>-2.71e-01</td>
<td>0.00e+00</td>
<td>2.60e-01</td>
<td>6.54e+00</td>
</tr>
<tr>
<td>max</td>
<td>1.35e-01</td>
<td>1.76e+00</td>
<td>-8.95e+00</td>
<td>-5.74e-01</td>
<td>3.75e-02</td>
<td>8.68e-01</td>
<td>1.12e+01</td>
</tr>
<tr>
<td>min</td>
<td>4.18e-02</td>
<td>2.98e-01</td>
<td>-2.59e+00</td>
<td>-7.10e-02</td>
<td>1.65e-02</td>
<td>7.66e-02</td>
<td>3.84e+00</td>
</tr>
</tbody>
</table>

Table 2.12: Statistics for the slope of a fitted linear regression for average, min, and max precipitation time series data
Figure 2.20: Geographical locations of projects with precipitation time series data whose regression models have outliers for intercept or slope values
Figure 2.21: Histograms for intercept and slope values without outliers for regression models of precipitation time series data
2.10 Project Starting Year

The starting year of projects are shown in figure 2.22, which shows that new projects are launching every year. The potential outcome framework assumes a binary treatment that is performed at some point in time such that one can distinguish a pre- and post-treatment situation. In this context, we consider the starting year of a project as the beginning of a treatment. To estimate the causal effect, we only consider pre-treatment data of covariates. So for each project, we use time series data before the projects started. In this dataset, the projects started between the years 2001 and 2012.

![Figure 2.22: Starting years of World Bank projects](image)

2.11 Project Funding

In addition to the project name, the World Bank provided information on the amount of funding for each project and the year it was implemented, alongside a number of other ancillary variables. A single project may take place at a number of locations, for example, if several schools are build for a project to improve education in a region or if a project improves the infrastructure of a region by building roads or improves an electrical grid.
The database also provides information on the number of locations associated with each project. These range from $n = 1$ to $n = 649$ project locations for a single project.

![Projects funding](image1)

![Locations of outliers in funding](image2)

**Figure 2.23:** Boxplot of funding amounts for projects and the geographical locations of outliers

<table>
<thead>
<tr>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.39e+08</td>
<td>1.93e+08</td>
<td>1.22e+06</td>
<td>3.51e+07</td>
<td>9.13e+07</td>
<td>1.88e+08</td>
<td>3.82e+09</td>
</tr>
</tbody>
</table>

**Table 2.13:** Statistics on the amounts of funding of World Bank projects (in US$)

In table 2.13, we can observe that median funding is about 9 million dollars. In figure 2.23 (b), projects located in Asia, Africa and South America have been invested in the most. A World Bank project may contain several sub-projects, as one large project may have multiple small projects in different places. In the World Bank data, we have the total commitment for the overall project (For World Bank projects, one large project many contains several subprojects in different locations) as shown in figure 2.23.
2.12 Data Characteristics

For variables listed in table 2.1, the temporal coverage is highly disparate. The temporal coverage of the covariates is variable across sources. For NDVI, precipitation, and temperature, we have highly granular, yearly information on characteristics at each World Bank project location. From this information, we generated additional information regarding the trend (positive or negative) before and after project implementation, as well as simple averages in the pre and post periods. As the project took various time lengths to finish, and because we do not have the real finished data of each project, we assume treatment is in effect after funds were distributed to the project. Many variables only have a single measurements; population density, accessibility to urban areas, slope, and elevation are all measured circa 2000, while distances to roads and rivers are measured circa 2010. We also investigated the correlation among covariates in figure 2.24.

For the two assumptions, we included no hidden confounders and overlap between treated and control on covariates. For the first assumption, we included all the covariates that may impact treatment and outcome to the best of our knowledge, as the world is complex and some confounders may exist that we are not aware of. For the second assumption, we check the overlap of all the covariates in the data set. As shown in figure 2.25, there exists enough overlap between the treatment and control for each covariate in the dataset.
Figure 2.24: Correlation among all covariates
Figure 2.25: Covariates overlap between treatment and control
2.13 Data Interpretation

One key attribute of causal attribution is a dataset that distinguishes between treated and untreated cases. In the case of a clinical trial, human beings who receive treatment might be contrasted to a control group of other humans of similar characteristics who do not receive a treatment. Because World Bank projects either exist or not, here we attempt to replicate the treated and untreated conditions by contrasting World Bank projects that were funded at very low levels (“control”) to those that were funded at high levels (“treated”). This is reflective of a hypothesis that the observed treatment effect should positively correlate with the amount of funding, i.e., huge amounts of funding are expected to have a bigger effect than small amounts of funding. Following this, we assign $W_i = 1$ if a project’s funding is in the upper third of all funded projects.

As a single project typically takes place at several project locations, we consider each project location as an individual unit, i.e., a school may be effective in one location, but not another, even if they were implemented by the same funding mechanism. Further, to capture potential geographic heterogeneity this might introduce, for each unit feature vector (i.e., selected covariates), we include the longitude and latitude of the project location. The total length of the feature vector is $d = 37$. All covariates are numerical, and their values are not normalized. For our outcome measure (i.e., the variable we seek to estimate the impact on), we contrast the pre-treatment and post-treatment average NDVI values at a project’s location. Let $ndvi_i(92,03)$ denote the average of NDVI values observed for project location $i$ over the years from 1992 to 2003 (the year before the project is implemented, which varies across projects; 2003 is used here for illustration). Let $ndvi_i(05,12)$ describe the corresponding value for the eight years after the project starts. The response $Y_i = ndvi_i(05,12) - ndvi_i(92,03)$ is thus the difference of the two averages. In order to calculate $Y^*$ for $Y$, we calculate the propensity score $e(x)$, which describes the expected likelihood
of treatment $W_i$ for a given unit of observation. As described above, while there are many methods for estimating $e(x)$, here we use logistic regression to provide a better comparison with the econometric approaches commonly employed in the international aid community.
<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean</th>
<th>Variance</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI average before project intercept</td>
<td>2.10e-01</td>
<td>0.00e+00</td>
<td>2.10e-01</td>
<td>2.90e-01</td>
</tr>
<tr>
<td>NDVI average before project slope</td>
<td>0.00e+00</td>
<td>0.00e+00</td>
<td>-1.00e-01</td>
<td>8.00e-01</td>
</tr>
<tr>
<td>average of max NDVI before project</td>
<td>2.20e-01</td>
<td>0.00e+00</td>
<td>9.00e-02</td>
<td>1.90e-01</td>
</tr>
<tr>
<td>latitude</td>
<td>1.77e+01</td>
<td>3.81e+02</td>
<td>-5.40e-01</td>
<td>-2.70e-01</td>
</tr>
<tr>
<td>longitude</td>
<td>4.05e+01</td>
<td>3.81e+03</td>
<td>-6.00e-01</td>
<td>-3.60e-01</td>
</tr>
<tr>
<td>distance to commercial rivers</td>
<td>1.93e+05</td>
<td>4.91e+12</td>
<td>1.96e+01</td>
<td>4.35e+02</td>
</tr>
<tr>
<td>distance to any rivers</td>
<td>3.51e+04</td>
<td>4.39e+11</td>
<td>2.35e+01</td>
<td>5.61e+02</td>
</tr>
<tr>
<td>distance to roads</td>
<td>3.01e+04</td>
<td>1.28e+11</td>
<td>1.46e+01</td>
<td>2.29e+02</td>
</tr>
<tr>
<td>travel time to cities</td>
<td>1.31e+02</td>
<td>6.36e+04</td>
<td>9.21e+00</td>
<td>1.27e+02</td>
</tr>
<tr>
<td>population 1990</td>
<td>7.62e+02</td>
<td>5.78e+05</td>
<td>9.28e+00</td>
<td>1.43e+02</td>
</tr>
<tr>
<td>population 1995</td>
<td>8.48e+02</td>
<td>7.19e+06</td>
<td>9.26e+00</td>
<td>1.38e+02</td>
</tr>
<tr>
<td>population 2000</td>
<td>9.29e+02</td>
<td>8.96e+06</td>
<td>9.45e+00</td>
<td>1.40e+02</td>
</tr>
<tr>
<td>slope</td>
<td>2.49e+00</td>
<td>1.65e+01</td>
<td>3.31e+00</td>
<td>1.49e+01</td>
</tr>
<tr>
<td>elevation</td>
<td>6.00e+02</td>
<td>8.00e+02</td>
<td>2.03e+00</td>
<td>7.57e+00</td>
</tr>
<tr>
<td>temperature average before intercept</td>
<td>1.75e+01</td>
<td>3.17e+01</td>
<td>-6.10e-01</td>
<td>-3.30e-01</td>
</tr>
<tr>
<td>temperature average before slope</td>
<td>3.00e-02</td>
<td>0.00e+00</td>
<td>3.50e-01</td>
<td>4.40e-01</td>
</tr>
<tr>
<td>temperature max before intercept</td>
<td>2.14e+01</td>
<td>3.82e+01</td>
<td>-2.70e+00</td>
<td>1.06e+01</td>
</tr>
<tr>
<td>temperature max before slope</td>
<td>3.00e-02</td>
<td>0.00e+00</td>
<td>6.50e-01</td>
<td>9.00e-01</td>
</tr>
<tr>
<td>temperature min before intercept</td>
<td>1.00e+01</td>
<td>1.15e+02</td>
<td>5.10e-01</td>
<td>2.93e+00</td>
</tr>
<tr>
<td>temperature min before slope</td>
<td>5.00e-02</td>
<td>5.00e-02</td>
<td>9.80e+00</td>
<td>1.46e+02</td>
</tr>
<tr>
<td>average of mean temperature before project</td>
<td>1.73e+01</td>
<td>2.93e+01</td>
<td>-5.50e-01</td>
<td>-3.40e-01</td>
</tr>
<tr>
<td>precipitation average before intercept</td>
<td>9.46e+01</td>
<td>3.12e+03</td>
<td>1.96e+00</td>
<td>7.81e+00</td>
</tr>
<tr>
<td>precipitation average before slope</td>
<td>5.00e-02</td>
<td>9.20e-01</td>
<td>1.36e+00</td>
<td>7.49e+00</td>
</tr>
<tr>
<td>precipitation max before intercept</td>
<td>1.72e+02</td>
<td>1.09e+04</td>
<td>7.70e-01</td>
<td>1.12e+00</td>
</tr>
<tr>
<td>precipitation max before slope</td>
<td>1.40e-01</td>
<td>3.10e+00</td>
<td>3.70e-01</td>
<td>3.75e+00</td>
</tr>
<tr>
<td>precipitation min before intercept</td>
<td>2.24e+01</td>
<td>3.12e+02</td>
<td>2.14e+00</td>
<td>5.57e+00</td>
</tr>
<tr>
<td>precipitation min before slope</td>
<td>4.00e-02</td>
<td>9.00e-02</td>
<td>2.39e+00</td>
<td>1.67e+01</td>
</tr>
<tr>
<td>average of mean precipitation before project</td>
<td>8.51e+01</td>
<td>1.52e+3</td>
<td>4.6e-01</td>
<td>-3.80e-01</td>
</tr>
<tr>
<td>project starting year</td>
<td>2.01e+03</td>
<td>1.10e+01</td>
<td>2.00e-02</td>
<td>-1.12e+00</td>
</tr>
<tr>
<td>total fund</td>
<td>1.39e+08</td>
<td>3.72e+16</td>
<td>6.36e+00</td>
<td>7.59e+01</td>
</tr>
<tr>
<td>lights average before project intercept</td>
<td>1.73e+01</td>
<td>1.19e+03</td>
<td>1.74e+00</td>
<td>3.78e+00</td>
</tr>
<tr>
<td>lights average before project slope</td>
<td>3.40e-01</td>
<td>1.02e+00</td>
<td>2.84e+00</td>
<td>1.24e+01</td>
</tr>
<tr>
<td>night luminosity trend before project</td>
<td>3.70e-01</td>
<td>8.90e-01</td>
<td>2.89e+00</td>
<td>1.14e+01</td>
</tr>
<tr>
<td>average of night luminosity before projects</td>
<td>1.52e+01</td>
<td>4.28e+02</td>
<td>1.24e+00</td>
<td>6.00e-02</td>
</tr>
</tbody>
</table>

Table 2.14: Moments
<table>
<thead>
<tr>
<th>Covariate</th>
<th>Distribution</th>
<th>First parameter</th>
<th>Second parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI average before project intercept</td>
<td>Normal</td>
<td>mean 2.10e-01</td>
<td>sd 5.80e-02</td>
</tr>
<tr>
<td>NDVI average before project slope</td>
<td>Normal</td>
<td>mean 1.19e-03</td>
<td>sd 1.17e-03</td>
</tr>
<tr>
<td>average of max NDVI before project</td>
<td>Normal</td>
<td>mean 2.20e-01</td>
<td>sd 6.00e-02</td>
</tr>
<tr>
<td>latitude</td>
<td>Normal</td>
<td>mean 1.77e+18</td>
<td>sd 1.95e+01</td>
</tr>
<tr>
<td>longitude</td>
<td>Normal</td>
<td>mean 4.05e+01</td>
<td>sd 6.17e+01</td>
</tr>
<tr>
<td>distance to commercial rivers</td>
<td>Normal</td>
<td>mean 1.93e+05</td>
<td>sd 7.00e+05</td>
</tr>
<tr>
<td>distance to any rivers</td>
<td>Gamma</td>
<td>shape 2.80e-03</td>
<td>rate 7.97e-08</td>
</tr>
<tr>
<td>distance to roads</td>
<td>Gamma</td>
<td>shape 7.08e-03</td>
<td>rate 2.34e-07</td>
</tr>
<tr>
<td>travel time to cities</td>
<td>Gamma</td>
<td>shape 2.70e-01</td>
<td>rate 2.05e-03</td>
</tr>
<tr>
<td>population 1990</td>
<td>Gamma</td>
<td>shape 1.00e-01</td>
<td>rate 1.32e-04</td>
</tr>
<tr>
<td>population 1995</td>
<td>Gamma</td>
<td>shape 9.50e-02</td>
<td>rate 1.16e-04</td>
</tr>
<tr>
<td>population 2000</td>
<td>Gamma</td>
<td>shape 9.00e-02</td>
<td>rate 1.04e-04</td>
</tr>
<tr>
<td>slope</td>
<td>Gamma</td>
<td>shape 3.70e-01</td>
<td>rate 1.50e-01</td>
</tr>
<tr>
<td>elevation</td>
<td>Weibull</td>
<td>shape 6.62e-01</td>
<td>scale 4.56e+02</td>
</tr>
<tr>
<td>temperature average before intercept</td>
<td>Normal</td>
<td>mean 1.75e+01</td>
<td>sd 5.63e+00</td>
</tr>
<tr>
<td>temperature average before slope</td>
<td>Normal</td>
<td>mean 3.00e-02</td>
<td>sd 2.00e-02</td>
</tr>
<tr>
<td>temperature max before intercept</td>
<td>Logistic</td>
<td>location 2.25e+01</td>
<td>scale 2.68e+00</td>
</tr>
<tr>
<td>temperature max before slope</td>
<td>Normal</td>
<td>mean 2.70e-02</td>
<td>sd 2.70e-02</td>
</tr>
<tr>
<td>temperature min before intercept</td>
<td>Logistic</td>
<td>location 1.01e+01</td>
<td>scale 6.01e+00</td>
</tr>
<tr>
<td>temperature min before slope</td>
<td>Normal</td>
<td>mean 5.00e-02</td>
<td>sd 2.20e-01</td>
</tr>
<tr>
<td>average of mean temperature before project</td>
<td>Normal</td>
<td>mean 1.72e+01</td>
<td>sd 5.43e+00</td>
</tr>
<tr>
<td>precipitation average before intercept</td>
<td>Gamma</td>
<td>shape 2.87e+00</td>
<td>rate 3.00e-02</td>
</tr>
<tr>
<td>precipitation average before slope</td>
<td>Logistic</td>
<td>location -1.00e-02</td>
<td>scale 4.20e-01</td>
</tr>
<tr>
<td>precipitation max before intercept</td>
<td>Gamma</td>
<td>shape 2.72e+00</td>
<td>rate 2.00e-02</td>
</tr>
<tr>
<td>precipitation max before slope</td>
<td>Logistic</td>
<td>location 9.00e-02</td>
<td>scale 8.70e-01</td>
</tr>
<tr>
<td>precipitation min before intercept</td>
<td>Gamma</td>
<td>shape 1.61e+00</td>
<td>rate 7.00e-02</td>
</tr>
<tr>
<td>precipitation min before slope</td>
<td>Logistic</td>
<td>location 2.00e-02</td>
<td>scale 1.20e-01</td>
</tr>
<tr>
<td>average of mean precipitation before project</td>
<td>Gamma</td>
<td>shape 4.76e+00</td>
<td>rate 6.00e-02</td>
</tr>
<tr>
<td>project starting year</td>
<td>Normal</td>
<td>mean 2.00e+03</td>
<td>sd 3.31e+00</td>
</tr>
<tr>
<td>total fund</td>
<td>Gamma</td>
<td>shape 5.21e-01</td>
<td>rate 3.74e-09</td>
</tr>
<tr>
<td>lights average before project intercept</td>
<td>Logistic</td>
<td>location 1.17e+01</td>
<td>scale 1.71e+01</td>
</tr>
<tr>
<td>lights average before project slope</td>
<td>Logistic</td>
<td>location 1.70e+01</td>
<td>scale 3.90e-01</td>
</tr>
<tr>
<td>night luminosity trend before project</td>
<td>Gamma</td>
<td>shape 1.60e-01</td>
<td>rate 4.20e-01</td>
</tr>
<tr>
<td>average of night luminosity before projects</td>
<td>Gamma</td>
<td>shape 5.30e-01</td>
<td>rate 4.00e-02</td>
</tr>
</tbody>
</table>

Table 2.15: Covariate distribution fitting
Chapter 3

Heterogeneous Causal Effect of World Bank Financed Projects

The World Bank provides billions of dollars in development finance to countries across the world every year\(^1\). As many projects are related to the environment, we want to understand the World Bank projects impact to forest cover. However, the global extent of these projects results in substantial heterogeneity in impacts due to geographic, cultural, and other factors. Recent research by Athey and Imbens has illustrated the potential for hybrid machine learning and causal inferential techniques which may be able to capture such heterogeneity. We apply their approach using a geolocated dataset of World Bank projects, and augment this data with satellite-retrieved characteristics of their geographic context (including temperature, precipitation, slope, distance to urban areas, and many others). We use this information in conjunction with causal tree (CT) and causal forest (CF) approaches to contrast ‘control’ and ‘treatment’ geographic locations to estimate the impact of World Bank projects on vegetative cover.

\(^1\)This chapter is taken almost verbatim from our paper published in [90].
3.1 Introduction

We frequently seek to test the effectiveness of targeted interventions - for example, a new website design or medical treatment. Here, we present a case study of using recent theoretical advances - specifically the use of tree-based analysis [5] - to estimate heterogeneous causal effects of global World Bank projects on forest cover over the last 30 years.

The World Bank is one of the largest contributors to development finance in the world, seeking to promote human well-being through a wide variety of programs and related institutions [1]. However, this goal is frequently at odds with environmental sustainability - building a road can necessitate the removal of trees; building a factory that supplies jobs can lead to the pollution of proximate forests. Multiple environmental safeguards have been put in place to offset these challenges, but relatively little is known about their efficacy across large scales.

We adopt the commonly applied approach of selecting “control” cases (i.e., areas where World Bank projects have very little funding) to contrast to “treated” cases (i.e., areas where World Bank projects have a large amount of funding). This is analogous to similar approaches in the medical literature, where humans are put into control and treatment groups, and individuals that are similar along all measurable attributes are contrasted to one another after a medicine is administered. This is necessary due to the generalized challenge of all observational studies: it is impossible observe the exact same unit of observation with and without a World Bank project simultaneously - in the same way it would be impossible to examine a patient that was and was not given medication at the same time. Further complicating the challenge presented in this paper is the scope of the World Bank - with tens of thousands of project locations worldwide, there is considerable variation in the aims of different projects, the project’s size, location, socio-economic, environmental, and historical settings. This variation makes traditional, aggregate estimates of impact
unhelpful, as such aggregates mask variation in where World Bank projects may be helping - or harming - the environment. Following this, we investigate the research question *What is the impact of world bank projects on forest cover?*

To examine this question, we first integrate information on the geographic location of World Bank projects with additional, satellite derived information on the geographic, environmental, and economic characteristics of each project. We apply four different models to this dataset, and contrast our findings to illustrate the various tradeoffs in these approaches. Specifically, we test Transformed Outcome Trees (TOTs), Causal Trees (CTs), Random Forest TOTs (RFTOTs), and Causal Forests (CFs). We follow the work of Athey and Imbens [5], who demonstrated how regression trees and random forests can be adjusted to estimate heterogeneous causal effects. This work is based on the Rubin Causal Model (or potential outcome framework), where causal effects are estimated through comparisons between observed outcomes and the “counterfactual” outcomes one would have observed under the absence of an aid project [41]. While traditional tree-based approaches rely on training with data with known outcomes, Athey and Imbens illustrated that one can estimate the conditional average treatment effect on a subset with regressions trees after an appropriate data transformation process. The convergence and consistency of trees and random forest have been studied in [51], [68],[69],[72],[83],[13], [18],[12],[15],[20], [21],[23], [32], [11] and [17]. Many approaches to estimating heterogeneous effects have emerged over the last decade. LASSO [77] and support vector machines (SVM) [81] may serve as two popular examples. For this paper, we focus on very recent tree-based techniques that are very promising for causal inference.

In [35], [38],[37],[65] ,[74], they used forest method to solve the heterogeneous causal effect problem.

Many research have been done in estimating heterogeneous treatment effects in [75],[24],[35],[61],[89],
In [74], Su et al. proposed a statistical test as the criterion for node splitting. In [5], Athey and Imbens derived TOTs and CTs, an idea that is followed up on by Wagner and Athey [82] with CF (causal forest, random forests of CTs), and similarly Denil et al. in [28] who use different data for the structure of the tree and the estimated value within each node. Random forests naturally gave rise to the question of confidence intervals for the estimates they deliver. Following this, Meinshausen introduced quantile regression forests in [50] to estimate a distribution of results, and Wagner et al [83] provided guidance for confidence intervals with random forests. Several authors, including Biau [10], recognize a gap between theoretical underpinnings and the practical applications of random forests.

The contribution of this paper is twofold: we evaluate and compare a number of proposed methods on simulated data where the ground truth is known and apply the most promising for the analysis of a real world data set. Practical experience results on tree-based causal inference methods are rare. To the best of our knowledge, this is the first investigation on the analysis of a spatial data set of world wide range with a large scale set of projects and dimensions. When it comes to applications for causal inference techniques, A/B testing for websites (such as eBay) is a more common [75]. A/B testing is conducted by diverting some percentage of traffic for a website A to a modified variant B of said website for evaluation purposes. This leads to a large amount of data with clearly defined treated and untreated groups where cases vary mainly by user activity. While the difference between A and B is precisely defined and typically small, the huge number of cases helps to recognize treatment effects. This is very different to the World Bank data which is both much more limited in size, and also spread all over the world (resulting in large diversity across projects). The rest of the paper is structured as follows. In Section 2, we present the basic methodology for the calculation of CT and CF. Section 3 introduces the data set, its
characteristics, preprocessing steps and the calculation of propensity scores necessary for the estimation of each type of tree. In Section 4, we present the outcome of the analysis. We conclude in Section 5.

3.2 Methodology

Causal inference is to a vast part a missing data problem as we can not observe a unit at the same time receiving and not receiving treatment to compare the outcomes. We introduce some notation and recall common concepts to be able to address this problem in a more formal way.

3.2.1 Causal Effects

In [87],[39],[49],[46], they analyze the use of propensity score. Suppose we have a data set with \( n \) independently and identically distributed (iid) units \( U_i = (X_i, Y_i) \) with \( i = 1, \cdots, n \). Each unit has an observed feature vector \( X_i \in \mathbb{R}^d \), a response (i.e., the outcome of interest) \( Y_i \in \mathbb{R} \) and binary treatment indicator \( W_i \in \{0,1\} \). For a unit-level causal effect, the Rubin causal model considers the treatment effect on unit \( i \) being \( \tau(X_i) = Y_i(1) - Y_i(0) \), the difference between treated \( Y_i(1) \) and untreated \( Y_i(0) \) outcome. One can be interested in an overall average treatment effect across all units \( U \) or investigate treatment effects of subsets that are characterized by their features \( X \). The latter describes heterogeneous causal effects and is often of particular interest. In our case, it is interesting to identify characteristics of subsets of projects where the environment is affected strongly (positive or negative) by a World Bank project. The heterogeneous causal effect is defined as \( \hat{\tau}(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] \) following [39].
3.2.2 Causal Tree

A regression tree defines a partition of a set of units $U_i = (X_i, Y_i)$ as each leaf node holds a subset of units satisfying conditions on $X$ expressed along the path from root node to leaf. This helps for the condition in $\hat{\tau}(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$. In observational studies, a unit is either treated or not, so we know either $Y_i(1)$ or $Y_i(0)$, but not both. However, one can still estimate $\tau(x)$ if one assumes unconfoundedness: $W_i \perp \perp (Y_i(1), Y_i(0)) \mid X_i$. Athey and Imbens [5] showed that one can estimate the causal effect as:

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

where $e(X_i)$ is the propensity score of project $i$ which is calculated by logistic regression, $T$ represents treatment units, and $C$ control units. Hence one can adapt the calculation of a regression tree to support calculation of $\hat{\tau}(X_i)$ by (4.2) by adjusting the splitting rule in the tree generation process.

In a classic regression tree, mean square error (MSE) is often used as the criterion for node splitting, and the average value within the node is used as the estimator. Following Athey and Imbens [5], we use (4.2) as the estimator and the following equation as the new MSE for any given node $J$ in the causal tree.

$$MSE = \sum_{i \in J} (Y_i(1) - Y_i(0) - \hat{\tau}(X_i))^2 = \sum_{i \in J} \tau(X_i)^2 - \sum_{i \in J} \hat{\tau}(X_i)^2$$

The right equation follows if one assumes that $\sum_{i \in J} \tau(X_i) = \sum_{i \in J} \hat{\tau}(X_i)$. The key observation is that $\sum_{i \in J} \tau(X_i)^2$ is constant and does not impact $\Delta MSE$. For a split, data in node $P$ is split into a left $L$ and right $R$ node, $\Delta MSE = MSE_P - MSE_L - MSE_R = \sum_{i \in P} \hat{\tau}(X_i)^2 - \sum_{i \in L} \hat{\tau}(X_i)^2 - \sum_{i \in R} \hat{\tau}(X_i)^2$. The ground truth $\tau(X_i)$ cancels out in $\Delta MSE$ and we can grow the tree without knowledge of $\tau(X_i)$. However, there is one more constraint
we need to add to the splitting rule aside from $MSE$. To use (4.2) for the calculation of $\hat{\tau}(X_i)$, neither set $T$ nor $C$ can be empty. Due to characteristics of the data in our applied study, we found that cases where only $C$ or $T$ units existed in children naturally emerged, so we added a corresponding additional stopping criterion to the splitting rule to prevent splits that would lead to situations where $T$ or $C$ had less than a fixed minimum cardinality.

### 3.2.3 Causal Forest

While a single causal tree allows us to estimate the causal effect, it leads to the problem of overfitting and subsequent challenges for pruning the tree. A common solution is to use an ensemble method such as bootstrap aggregating or bagging, namely a variant of Breiman’s random forest [16]. If one applies the random forest approach to causal trees, the result is called a causal forest. Computation of a causal forest scales well as it can naturally be run in parallel. The same adjustments for generating a single CT apply to the generation of a random forest of CTs. We implemented a causal forest algorithm with the help of the scikit learn package. We can estimate the causal effect $\tau_{CF}(X_i)$ from a causal forest (a set $CF$ of causal trees) for a unit $i$ as the average across the estimates obtained from its trees:

$$\hat{\tau}_{CF}(X_i) = \frac{1}{|CF|} \sum_{t \in CF} \hat{\tau}_t(X_i).$$

### 3.3 Experiments and Results

We follow a two stage procedure to examine the effectiveness of both the CT and CF algorithms, specifically considering our unique context of the effectiveness of World Bank projects. First, we test and evaluate which approach is most suited to our application using simulated synthetic data where we know the ground truth and where we can vary the size of sample data. Second, we apply these algorithms to examine the efficacy of World Bank projects based on satellite imagery. We implemented the CT and CF algorithms as
well as Athey and Imbens transformed outcome tree (TOT) approach [5] and a random forest variant of TOT (RFTOT) using scikit-learn. The latter serve as a baseline for the performance of CT and CF algorithms.

3.3.1 Experimental Results for Simulated Data

First, we iteratively simulate synthetic datasets with known parameters to evaluate how the estimation of propensity score, dataset size, and degree of similarity between the control and treatment groups impact the accuracy of the result. To do this, we follow a bi-partite data generation process, in which two equations are used (one for treated cases and another for control cases).

We use each of the following two equations to produce one half of all data points. $Y_i^1$ gives the result for treated cases; $Y_i^0$ is for the control group. Here, from $x_1$ to $x_8$, $x_j \sim \mathcal{N}(0, 1)$ as well as $\varepsilon \sim \mathcal{N}(0, 1)$.

$$Y_i(1) = W_i^1 + \sum_{j=1}^{k} x_j \ast W_i^1 + \sum_{j=1}^{8} x_j + \varepsilon, \quad Y_i(0) = W_i^0 + \sum_{j=1}^{k} x_j \ast W_i^0 + \sum_{j=1}^{8} x_j + \varepsilon \quad (3.3)$$

As used in Table 3.1, $k$ is defined as the number of covariates which contribute to heterogene-

![Figure 3.1](image-url)

**Figure 3.1**: Estimated treatment effects for randomized assignment, $e(x) = 0.5$
ity in the causal effect. The true value of the causal effect is then \( \tau(X_i) = Y_i(1) - Y_i(0) = 1 + \sum_{i=1}^{k} x_i \), with \( W^1 = 1 \) and \( W^0 = 0 \).

The first scenario we examine considers synthetic datasets with a randomized treatment assignment (each unit has the same probability to be treated, \( c(x) = 0.5 \)). Figure 3.1 shows corresponding results for \( n=2000 \), and includes both single tree and random forest implementations of Transformed Outcome Trees (TOT; [5]) for comparison. The resultant distributions all encompass the true mean results, but with considerable difference in overall metrics of error. The Causal Forest approach is the most accurate across all simulations as well as the tightest overall distribution; this is in contrast to the TOT forest implementation. For single trees, the CT performs much better than the TOT and even outperforms the RFTOT.

<table>
<thead>
<tr>
<th>sample size</th>
<th>CF mean</th>
<th>CF std</th>
<th>CT mean</th>
<th>CT std</th>
<th>TOT mean</th>
<th>TOT std</th>
<th>RFTOT mean</th>
<th>RFTOT std</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.60</td>
<td>0.001</td>
<td>1.27</td>
<td>0.02</td>
<td>9.96</td>
<td>0.24</td>
<td>7.74</td>
<td>0.13</td>
</tr>
<tr>
<td>5000</td>
<td>0.58</td>
<td>0.001</td>
<td>0.99</td>
<td>0.02</td>
<td>7.95</td>
<td>0.03</td>
<td>5.61</td>
<td>0.05</td>
</tr>
<tr>
<td>10000</td>
<td>0.51</td>
<td>0.00001</td>
<td>0.86</td>
<td>0.005</td>
<td>7.45</td>
<td>0.02</td>
<td>5.14</td>
<td>0.02</td>
</tr>
</tbody>
</table>

**Table 3.1**: Mean square error (forest has 1000 trees, feature ratio = 0.8)

The second scenario considers synthetic datasets with varying numbers of observations (\( n=1000, 5000, \) and \( 10,000 \)). We calculate the mean square error for CT, CF, TOT and RFTOT. The results in Table 3.1 show that - as expected - the error gets smaller as the number of observations increases. Of particular importance, we note that in the case of smaller datasets, the CF implementation strongly outperforms the single-tree CT implementation under all the scenarios we test.

We also test the convergence of each method as the size of data increases, as shown in Figure 3.2. Figure 3.2a shows the MSE of each methods with increasing data size, while Figure 3.2b shows a zoomed-in version of the MSE of the CF approach (due to the lower
Figure 3.2: MSE changes with data size

magnitude of MSE observed). At least for this specific data generation process, the CF and CT outperform other approaches, which is why we focus on them for the analysis of the World Bank data set where we can not measure accuracy.

3.3.2 Results for World Bank Data

Following the simulation results, we seek to identify and contrast the benefits and drawbacks associated with applying CT and CF approaches to a real-world scenario. In this case study, we identify the impact of international aid - specifically, World Bank projects - on forest cover. First, we use a single CT to estimate the causal effect $\hat{\tau}(X_i)$ of a single project $i$ with (equation 4.2) applied to the leaf where the project is located. Second, we implement a Causal Forest.

While our simulations, as well as the existing literature, suggest the Causal Tree has many drawbacks relative to a Causal Forest, it can enable practitioners to make inferences that are precluded by forest-based approaches. Most notably, the structure of single trees can provide insight into the explicit drivers of impacts - in this case, of World Bank projects. As an example, in the Causal Tree implementation here, we find that the year a project
started was an important driver of effectiveness - specifically, projects starting before 2005 were more effective than those after 2005. This type of insight is particularly helpful, as it allows for analysis into the causes of impact heterogeneity. However, the lack of information on the robustness of findings in a single tree approach, coupled with the relative inaccuracy of CT as contrasts to CF, indicates that such findings should be approached with caution until better methods for identifying the robustness of CT tree shapes are derived.

Figure 3.3: CF calculated distributions of treatment effect estimates for specific projects: (a) Saint Lucia Hurricane Tomas Emergency Recovery Loan; (b) Sustainable Tourism Development Project; (c) Emergency Infrastructure Reconstruction Project.

The Causal Forest (CF) implementation represents a set of CTs and thus creates a distribution of values for each World Bank project $i$. These distributions are then aggregated to a single value to estimate $\hat{\tau}(X_i)$, or the distributions themselves are analyzed to examine the robustness of a given finding. In Figure 3.3, we show the detailed distributions for selected example projects. These examples provide an illustration of how applied CF results can provide indications not only of what projects are likely having a negative impact on the environment, but also the robustness of these estimates. By writing a second-stage algorithm which identified projects with distributions following certain characteristics (i.e., a mean centered around 0 with a Gaussian distribution; a negative-centered mean with a left-skewed distribution), it is possible to highlight the subset(s) of projects for which more robust findings exist. Figure 3.4 (a) shows a histogram of CF calculated $\hat{\tau}(X_i)$ values for
all world bank projects in our data set. Most of the projects have a slightly negative to no impact on the forest cover, which is in line with World Bank objectives to offset potential negative environmental outcomes. Figure 3.4 (b) provides evidence that while the World Bank is generally successful in meeting it’s goal of mitigating environmental impacts, the rate at which positive and negative deviation occurs is highly variable by geographic region. We can see that most outliers are in the positive direction, with Asia being a notable exception. The projects in Oceania are in a narrow range, however, projects in other continents have a wide range.

![Figure 3.4](image)

**Figure 3.4**: (a) Causal effect distribution of all World Bank projects combined and (b) separated by continents

While both the CT and CF approaches allow for the examination of the relative importance of factors in driving heterogeneity, the interpretation and robustness of these findings is highly variable. In the case of the CT, the position of a variable in the single tree can be interpreted as importance; i.e., splits higher in the tree are more influential on the results, and path-dependencies can be examined. However, the robustness of the shape of the CT approach is unknown, and both our simulations and existing literature suggest CT findings are likely to be less accurate than CF implementations. Conversely, in a CF each covariate can be ranked across all trees in terms of the purity improvements it can provide, giving a
relative indication of importance across all trees??. While these findings are more robust, they do not enable the interpretation of explicit thresholds (i.e., the year variable may be important, but the explicit year that is split on may change in the RF approach), and path dependencies are not made explicit. In our case study, we find that the first five variables in the CT and CF cases are stable between approaches, but we identify significant variance in deeper areas of the tree. For a practitioner, this allows an understanding of what the major drivers of aid effectiveness are; for example here, the purity metric highlights the dollars committed and environmental conditions as major drivers of forest cover loss, and also highlights a disparity between projects located at different latitudes; all factors which can enable a deeper understanding of what is causing success and failure in World Bank environmental initiatives. This is consistent with past findings which illustrate a stable set of covariates in the top-level of trees across a CF [75]. Further, we note that the 15 most highly ranked covariates in the CF approach are generally uncorrelated, providing an indication that the information they provide is not redundant??. However, we leave the interpretation of the shape of the random forest, and the insights that can be gained from it, to future research.

3.4 Discussion and Conclusions

This paper sought to examine the research question *What is the impact of world bank projects on forest cover?* To examine this, we contrasted four different approaches all based on variations of regression trees and random forests of trees: Transformed Outcome Trees (TOTs), Causal Trees (CTs), Random Forest TOTs (RFTOTs), and Causal Forests (CF). We found that the method selected can have significant influence on the causal effect (or lack thereof) estimated, and provide evidence suggesting CF is more accurate than alternatives in our study context. By applying the CF approach to the case of World Bank
projects, we were able to compute estimates for causal effects of individual projects; further, the prominent appearance of some covariates in trees provided us with guidance on which covariates were most important in mediating the impacts of World Bank projects. While - for most projects - the effect on forest cover is close to zero, we identified some notable exceptions, positive as well as negative ones. We also identified two key questions that have not yet been answered in the academic literature. The first of these is how to select proper limitations on the makeup of terminal nodes - i.e., if splits that result in nodes without both control and treatment cases should be prevented, omitted, or otherwise constrained. Even after propensity score adjustments, terminal nodes with no adequate comparison cases become difficult (if not impossible) to interpret. Second, there is little literature in the machine learning space regarding how to cope with spatial spillover between treated and control cases. The Stable Unit Treatment Value Assumption (SUTVA) is common practice, but in practice the effects of a project can not be expected to be purely local in nature when observations are geographically situated.
Chapter 4

Simulation Study in Quantifying Heterogeneous Causal Effects

Quantifying the impact of an intervention or treatment in a real setting is a common and challenging problem\(^1\). For example, we would like to calculate the environmental implications of aid projects in third world countries that target economic development. For causal inference problems of this kind, the Rubin causal model is one of several popular theoretical frameworks that comes with a set of algorithmic methods to quantify treatment effects. However, for a given data set, we neither know the ground truth nor can we easily increase the size of the data set. So, simulation is a natural choice to evaluate the applicability of a set of methods for a particular problem. In this paper, we report findings of a simulation study with four causal inference approaches, namely two single tree approaches (transformed outcome tree, causal tree), and two random forest versions of the former.

\(^1\)This chapter is taken almost verbatim from a paper we published in [91].
4.1 Introduction

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

The work presented here is based on the Rubin Causal Model (or potential outcome framework), where causal effects are estimated through comparisons between observed outcomes and the “counterfactual” outcomes one would have observed under the absence of an intervention [41]. In the causal inference literature, the terminology to describe this follows a medical point of view. The intervention is called a treatment and all observed units are separated into two subsets: a group of treated cases versus a group of control cases. The Rubin causal model is a common framework to model and evaluate causal inference. The outcome of such an analysis is either the average treatment effect observed on the whole population of units or the conditional average treatment effect one observes for a specific subset of units that share some particular characteristics. This leads to the need to describe units by relevant properties, i.e., values for a number of variables or features that are called covariates in this context. The interest in the conditional average treatment effect naturally arises from the fact that treatments are not prescribed randomly or in general but to address an observed condition or situation. For example, a third world aid project to fund a hospital in general is good but for sure it will have a bigger impact - a bigger conditional average treatment effect - if placed in an area that is currently underserved with medical treatment facilities and where people frequently suffer from diseases that are easily curable.

In observational studies, especially in the medical and social sciences, there is interest in the estimation of such heterogeneous causal effects.

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

Many approaches to estimating heterogeneous effects have emerged over the last decade.
LASSO [77] and support vector machines (SVM) [81] may serve as two popular examples,
however, we have limited this analysis to a small subset of techniques that are based on
regression trees. Specifically, we test Transformed Outcome Trees (TOTs), Causal Trees
(CTs), Random Forest TOTs (RFTOTs), and Causal Forests (CFs). We follow the work
of [5], who demonstrated how decision trees and random forests can be adjusted to esti-
mate heterogeneous causal effects. While traditional tree-based approaches rely on training
with data with known outcomes, Athey and Imbens illustrated that one can estimate the
conditional average treatment effect on a subset with regression trees after an appropriate
data transformation process. The historic lineage of causal inferential study using trees is
relatively young, but rapidly growing. In [74], Su et al. proposed a statistical test as the
criterion for node splitting. In [5], Athey and Imbens derived TOTs and CTs, an idea that
is followed up on by Wagner and Athey [82] with CF (causal forest, random forests of CTs),
and similarly Denil et al. in [28] who use different data for the structure of the tree and the
estimated value within each node. Random forests naturally gave rise to the question of
confidence intervals for the estimates they deliver. Following this, Meinshausen introduced
quantile regression forests in [50] to estimate a distribution of results, and Wagner et al [83]

64
provided guidance for confidence intervals with random forests. Several authors, including Biau [10], recognize a gap between theoretical underpinnings and the practical applications of random forests.

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

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

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

4.2.1 The Rubin Causal Model And Conditional Average Treatment Effects

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

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

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

\[
Y_i^* = Y_i \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))}, \tag{4.1}
\]

and the propensity score function \( e(x) \) is defined as \( e(x) = \mathbb{E}[W_i \mid X_i = x] \). In laboratory experiments as well as in simulation studies, it is common to randomly assign a treatment.
to a unit and to use \( e(x) = 0.5 \) to obtain balanced group sizes for treated and control groups. In reality, a certain disposition \( X_i \) will either qualify/demand for treatment such that a treatment will be assigned in most cases or disqualify for treatment and a treatment will be rarely assigned. The propensity score accounts for this effect. Several approaches to estimate the propensity score can be selected \([?\text{, }40]\); for the world bank data set, for instance, we use logistic regression in order to provide a stronger comparison to econometric modeling approaches most commonly employed by the international development community today.

### 4.2.2 Regression Tree For Causal Inference

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

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

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

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

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

### 4.2.3 Transformed Outcome Tree

A TOT is a regression tree that uses \( Y^* \) for \( f(x) \). As mentioned above, the transformed outcome is calculated with (4.1), then a traditional regression tree method is employed to generate a TOT for the given data set. The causal effect is subsequently estimated with 
\[
\tau(x) = \mathbb{E}[Y^* \mid X_i = x],
\]
where feature vector \( x \) leads to a leaf of the tree and \( \hat{\tau}(x) \) is the average transformed outcome of all units in the corresponding leaf of the tree.
4.2.4 Causal Tree

When a regression tree is used to estimate heterogeneous causal effects, tree construction is a key step. In a classic regression tree, mean square error (MSE) is often used as the criterion for node splitting, and the average value within the node is used as the estimator. Following Athey and Imbens [?], we use (4.2) as the estimator, and we replace the traditional MSE by summing $Y_i^* - \hat{\tau}(X_i)$.

$$\hat{\tau}(X_i) = \sum_{i \in T} Y_i \cdot \frac{W_i / \hat{e}(X_i)}{\sum_{j \in T} W_j / \hat{e}(X_j)} - \sum_{i \in C} Y_i \cdot \frac{(1 - W_i) / (1 - \hat{e}(X_i))}{\sum_{j \in C} (1 - W_j) / (1 - \hat{e}(X_j))}, \quad (4.2)$$

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

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

4.2.5 Random Forest

A random forest is a set of regression trees that are independently built using a bootstrap sample of the given data set and the splitting criterion applied in the construction only selects the best split among a subset of all possible predictors randomly chosen at that node [16]. The random forest concept can be applied to any approach that computes a single tree, as it is the case for the TOTs or CTs. To obtain an estimate for $\hat{\tau}(x)$ from a forest of trees, one calculates the average of all individual estimates $\hat{\tau}(x)$ that one obtains
from each single tree in the forest. The key idea is that the trees represent the variability in
the construction of trees due to the variability in the data (exposed by the bootstrapping
and random selection of a subset of candidate covariates for a split) and that the averaging
for the overall result creates a robust estimate for $\hat{\tau}(x)$ thanks to the large number of
independently generated trees.

We rely on the randomForest R package that implements Breiman’s approach and im-
ploved a causal forest algorithm with the help of the scikit learn package.

### 4.3 Simulation Model

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

$$Y_i(1) = W_i * (c + \sum_{i=1}^{k} \beta_i x_i) + \sum_{i=k+1}^{d} \beta_i x_i + \beta_0 + \varepsilon = c + \sum_{i=1}^{d} \beta_i x_i + \beta_0 + \varepsilon,$$  \hspace{1cm} (4.3)

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

For the control cases, we use

\[
Y_i(0) = W_i \cdot (c + \sum_{i=1}^{k} \beta_i x_i) + \sum_{i=k+1}^{d} \beta_i x_i + \beta_0 + \varepsilon = \sum_{i=k+1}^{d} \beta_i x_i + \beta_0 + \varepsilon, \tag{4.4}
\]

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

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

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

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

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

4.3.1 Configuring the Simulation Model to Approximate the Real Data Set

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

The existing data set has $d=37$ covariates and 16369 data points and the challenge is to recognize the difference in average NDVI values before and after a project starts as a treatment effect. Of course, the ground truth for the real data is not known. For the simulation study, we want to see if tree-based causal inference techniques apply to a data set of this kind. We choose $d=37$ for the simulation model and for each covariate in our data set, we perform a distribution fitting. Figure 4.1 shows an example for the variable that describes the average of annual maxima of NDVI values for all years before the start of the project with a q-q plot on the left and that shows to what extent the observations match with a Normal distribution. The right side of Figure 4.1 shows a heat map of the correlation among all covariates and correlations are moderate as the predominantly cooler colors indicate. The individual covariates show limited correlation such that we can sample individual entries for $(x_1, \ldots, x_d, \varepsilon)$ independently and from a distribution that we fitted for each covariate. For the error term $\varepsilon$ we decided to use a normal distribution $N(0, 1)$.

For the treatment effect in the simulation model, we need to choose a value for $k$ and select the covariates for $\tau_i = c + \sum_{i=1}^{k} x_i$. We follow the common 80/20 rule that suggests that the vast majority of an effect (about 80%) is caused by a small minority of parameters (about 20%). So, we consider different scenarios with $k = 1, 2, 4, \text{ and } 8$ variables and select a random subset $k$ covariates out of $d = 37$ for our experiment. In this way, we have a stochastic input model that we can use to derive data sets of any size with a known ground truth on treatment effect as well as the role of individual covariates that are responsible for the effect.
4.4 Experiments and Results

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

4.4.1 Accuracy of Causal Effect Estimate

Before we look into a realistic data set, we want to report on reassuring findings for a more common and simpler model that is closer to the ones typically analyzed in the literature.
We exclusively use standard normal distributions $\mathcal{N}(0,1)$ for all covariates and $\varepsilon$ which implies $\beta_i = 1$ for $i > 0$ and $\beta_0 = 0$ for Eqs. 4.3 and 4.4. We analyze a model configuration for $n = 2000$, $k = 1$, $d = 9$, and $c' = 1$. We use a random treatment assignment such that each unit has the same probability to be treated, $e(x) = 0.5$. On the left, Figure 4.2 shows box plots for the treatment effect $\tau(x)$ in column TRUTH as well as estimates $\hat{\tau}(x)$ for CF, CT, TOT and RFTOT in corresponding columns. The distribution of $\tau(x)$ shows that the treatment effect is heterogeneous. The resultant distributions all encompass the true mean results, but with considerable difference in overall metrics of error. In line with published results, the Causal Forest approach is the most accurate across all simulations, random forest variants perform better than single tree variants and the CT is better than the TOT. Of course, the accuracy is expected to depend on the amount of data that is available. So, we also test the convergence of each method as the size of data increases, as shown in the middle diagram in Figure 4.2. We measure the mean squared error (MSE) between estimated $\hat{\tau}(x)$ and the ground truth of $\tau(x) = c + \sum_{i=1}^{k} \beta_i x_i$ to evaluate the accuracy of the estimates. It shows the MSE of each method with increasing data size, while the diagram to the right shows a zoomed-in version of the MSE of the CF approach (due to the lower magnitude of MSE observed). As expected, accuracy improves with the amount of data, random forest approaches generally show a better performance than single tree approaches and CF shows best results. The box plots in the right diagram of Figure 4.2 result from repeated calculations and show the variability of MSE in response to the variability in inputs as well as the general convergence for increasing values of $n$.

Seeing positive results that are consistent with original research publications motivates us to move forward and analyze scenarios closer to our real data set. The first scenario we examine considers a simulated dataset with a randomized treatment assignment (each unit has the same probability to be treated, $e(x) = 0.5$). For our stochastic model, we used $k = 1$, 

75
Figure 4.2: Estimated treatment effects (left), MSE changes with data size (middle), and for CF (right).

\( d = 37, \ c' = 1, \) and \( n = 20000 \). As we increase the number of covariates \( d \), we report findings for two variants, one still samples values from a normal distribution \( \mathcal{N}(0, 1) \) independently for each covariate and a second one where all covariates are sampled independently from the distributions that we fitted to the real data set. So the second configuration models the situation where a single covariate is responsible for the treatment effect in a setting that resembles our real data set in terms of dimensions \( d = 37 \), covariate distributions, and size \( n = 20000 \). Note that \( E[\tau(x)] = 1 \) for the selected parameter settings and sampling distributions for covariates, in particular due to the settings of \( \beta_i \) coefficients and \( c' = 1 \).

We generate a set of data points for each experiment with about half for treated, half for the control group. We are interested in comparing the accuracy of the four different causal inference techniques. We measure the MSE between estimated \( \hat{\tau}(x) \) and the ground truth \( \tau(x) = c + \sum_{i=1}^{k} \beta_i x_i \) to evaluate the accuracy of the estimates. Note that this is not the artificial best case configuration, so the data set does not contain exactly matching pairs of treated and untreated units. All forests are computed with 1000 trees.

The box plots in Figure 4.3 show the distribution of \( \tau(x) \) values for the ground truth and estimates \( \hat{\tau}(x) \) for a causal forest CF and a causal tree CT for the first scenario. The box plot for the ground truth (TRUTH) shows that the true causal effect values vary only very lightly on the given scale, which is plausible as only a single random variable creates
some variability for it. While CF and CT are on the right scale, none of them have an interquartile range that matches well with the true distribution. CF notably underestimates the variability in $\tau(x)$. The box plots do not include results for TOT and RFTOT as their box plots spread wider by orders of magnitude and distort the visualization. This is easier to see in the numerical values in the table of Figure 4.3 which reports some basic statistics on MSE values (and not treatment effects) seen across a repeated set of 5 experiments for the scenario with sampling from a normal distribution on top and from the fitted distributions on the bottom. The sharp increase in $d$ made it more difficult to estimate $\tau(x)$ for the chosen value of $n$. For the normal distribution scenario, we can see that results for CF and CT are in a range that one could expect to achieve reasonably accurate results for some higher settings of $n$ if one takes trends into account that are seen on the right side of Fig. 4.2. TOT and RFTOT perform one to two orders of magnitude worse than CF and CT, which is consistent with what we saw in the previous experiment (graph in the middle of Fig. 4.2). Using fitted distributions instead of normal distributions makes these issues even more pronounced. MSE values for CF and CT improve for fitted distributions compared to normal distributions but get worse for TOT and RFTOT. We attribute the improvement

<table>
<thead>
<tr>
<th>$\mathcal{N}(0, 1)$</th>
<th>CF</th>
<th>CT</th>
<th>TOT</th>
<th>RFTOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>1.70E0</td>
<td>2.56E0</td>
<td>1.28E2</td>
<td>4.15E1</td>
</tr>
<tr>
<td>max</td>
<td>1.76E0</td>
<td>2.76E0</td>
<td>1.31E2</td>
<td>4.27E1</td>
</tr>
<tr>
<td>mean</td>
<td>1.72E0</td>
<td>2.68E0</td>
<td>1.29E2</td>
<td>4.20E1</td>
</tr>
<tr>
<td>var</td>
<td>6.0E-4</td>
<td>6.0E-3</td>
<td>2.43E0</td>
<td>2.5E-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>fitted</th>
<th>CF</th>
<th>CT</th>
<th>TOT</th>
<th>RFTOT</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>5.7E-2</td>
<td>2.9E-1</td>
<td>5.13E2</td>
<td>1.71E2</td>
</tr>
<tr>
<td>max</td>
<td>6.3E-2</td>
<td>3.3E-1</td>
<td>5.46E2</td>
<td>1.82E2</td>
</tr>
<tr>
<td>mean</td>
<td>6.9E-2</td>
<td>3.2E-1</td>
<td>5.22E2</td>
<td>1.74E2</td>
</tr>
<tr>
<td>var</td>
<td>5.4E-6</td>
<td>1.9E-4</td>
<td>1.8E2</td>
<td>2.1E1</td>
</tr>
</tbody>
</table>

**Figure 4.3:** Estimated treatment effect distributions for a single experiment with sampling from fitted distributions, MSE statistics for repeated runs for covariates sampled from normal and fitted distributions.
for CF and CT in part to the feature scaling that limits the range of covariates to $[0, 1]$ such that $\tau(x) \geq 1$. As TOT and RFTOT are clearly underperforming, we focus our analysis on CF and CT for the following.

Looking at a single covariate to determine the causal effect is a particular corner case. Figure 4.4 shows results for the same model configuration and for a single experiment for increasing values of $k = 1, 2, 4, \text{and } 8$. The values in the table show that MSE values tend to remain in the same order of magnitude. For both distribution scenarios, values increase with increasing values of $k$. This means when there are more confounders in the model, it is harder to estimate the causal effect. The box plots in Figure 4.4 show causal effect values for the case of sampling from fitted distributions and $k = 8$ and essentially match with what we have seen in Fig. 4.3 before, but for a ground truth that has more variability that is not adequately matched by CF or CT results. However, CF consistently achieves better MSE values than CT. For the scenario with fitted distributions, the rightmost columns in the table in Fig. 4.4 show small error values as before, which seem less affected by $k$ than corresponding results for the Normal distribution case.

The estimate of $\hat{\tau}(x)$ is based on the comparison of specific subsets of treated and untreated units. The accuracy of the outcome is influenced by the quality of the data set which can provide treated and untreated units that are more or less comparable or similar.
to each other. In order to see what CT and CF can achieve if applied to high quality data, we provide an artificial best case. We repeat the experiments with the same parameter settings and ranges but produce exact pairs of treated and untreated units. Of course, this is an impossible best case where the causal treatment effect calculation is trivial, but the question is if CT and CF approaches benefit from this. The question behind this is to what extent the binning that the tree calculation performs in its leaf nodes introduces errors into this causal effect calculation. Figure 4.5 shows corresponding results that we can directly compare with Figure 4.4. For CF, the box plot with $\hat{\tau}(x)$ estimates and the MSE values for the $\mathcal{N}(0,1)$ case do not change significantly. When sampling from the fitted distributions, we see a moderate reduction in the MSE values for CF but a significant reduction for CT that makes CT even more accurate than CF (with the exception of $k = 8$). The box plot shows that CT aligns well with the center of the true distribution but underestimates the interquartile range and variance. The CT approach benefits from this best case scenario. The MSE values for CT are slightly better than the ones for CF for both sampling distributions (with the exception of $k = 8$ for the fitted case). We see two main reasons for the CF to essentially retain its MSE values: the generation of a causal forest considers random subsets of values, which can break up the perfect pairs of data points we provide. Secondly, having a minimum set of treated and untreated cases in each

<table>
<thead>
<tr>
<th>k</th>
<th>$\mathcal{N}(0,1)$</th>
<th>fitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.63 0.99 0.03</td>
<td>0.007</td>
</tr>
<tr>
<td>2</td>
<td>2.57 2.00 0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>4.42 3.96 0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>8</td>
<td>8.16 7.91 0.07</td>
<td>0.08</td>
</tr>
</tbody>
</table>

\textbf{Figure 4.5:} Estimated treatment effects for CT and CF for $k = 8$ and MSE values for varying values of $k$ for the artificial best case scenario with matching pairs of data points for treated and untreated cases.
leaf node will imply that even if perfectly matching pairs are present in a leaf node, the treatment effect is calculated with average values that may also include data points without a matching unit. So, if we only subsample the treated units, this achieves better results.

Comparing results in Figs. 4.4 and 4.5 shows that a high quality data set with exactly matching pairs of treated and untreated cases does not lead to perfect results being calculated by the CF. We find that the required minimum leaf size is a parameter that also influences the quality of the estimated treatment effect. For the artificial best case, a minimum leaf size of two is a promising candidate. So, we exercise these experiments again but this time adjust the parameter for the splitting rule to allow for a single treated and a single untreated case in a leaf node. The latter case is denoted as $Min = 2$ in Table 4.1, while columns under $Min = 10$ denote that the splitting rule requires a total of 10 elements in a leaf node with at least 1 of each kind (treated or untreated). The results for the $\mathcal{N}(0,1)$ case show that CT can get exact results for a leaf size of 2 in the artificial best case. However, there are opposite trends: if one moves from the best case scenario to unpaired random samples, i.e. one reduces the quality of the data set, then increasing the minimum leaf size from two to ten helps the CT approach while for the best case scenario on its own, an increase in the leaf size increases MSE values. For the $\mathcal{N}(0,1)$ case, the CF approach is rather insensitive to changes of the minimum leaf size which suggests smaller value settings to include the corner case of the best case scenario.

If we sample data points from distributions fitted to the world bank data, we see that the CT approach for the artificial best case and the minimum leaf size of two is accurate. An increase of the leaf size for the best case scenario introduces errors for CT. However, for the regular case of unpaired random samples, a minimum leaf size of ten slightly reduces MSE for CT. For the CF approach, one can see that a minimum leaf size of ten in general is better than two regardless of the presence of perfectly matching data points or not.
Table 4.1: MSE results for different configurations of the splitting rule.

<table>
<thead>
<tr>
<th></th>
<th>Artificial Best Case</th>
<th>Unpaired Random Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N(0,1) Min=2</td>
<td>Min=10</td>
</tr>
<tr>
<td>k</td>
<td>CT</td>
<td>CF</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>1.24</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>1.83</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>2.87</td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>4.58</td>
</tr>
<tr>
<td>fitted</td>
<td>Min=2</td>
<td>Min=10</td>
</tr>
<tr>
<td>k</td>
<td>CT</td>
<td>CF</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>0.13</td>
</tr>
<tr>
<td>8</td>
<td>0.00</td>
<td>0.18</td>
</tr>
</tbody>
</table>

4.4.2 Convergence Rate for Increasing Data Size

Of course, one can not expect to find perfectly matching pairs in the data set as for our artificial best case scenario. It is more realistic to assume that with an increasing size of the data set, we will see more units that are similar in their covariate values and thus expect that the quality of the approximation of \( \tau(x) \) improves with \( n \), the size of the data set. Following this, we also test the convergence of the causal inference methods as the size of data increases, as shown in Figure 4.6.

We consider a model configuration with \( k = 1, d = 37, c' = 1 \) for increasing values of \( n \). The minimum leaf size is 10. We sample covariates from fitted distributions. This is not the artificial best case scenario. This experiment essentially tries to achieve the observed decrease in MSE values for increasing \( n \) that we have seen for the lower dimensional case in Fig. 4.2 for a model configuration with more covariates and sampling from different distributions. Figure 4.6 shows the MSE of CT and CF for a series of experiments with increasing data size that starts with \( n = 2000 \) and increases \( n \) by more than one order
of magnitude. The observed results do show a pronounced underlying trend and confirm observations for the lower dimensional case.

### 4.5 Discussion and Conclusions

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

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

There are a number of issues: at a very technical level for the splitting rule of trees, what is the best way to select proper limitations on the makeup of terminal nodes - i.e., if
splits that result in nodes without both control and treatment cases should be prevented, omitted, or otherwise constrained. Even after propensity score adjustments, terminal nodes with no adequate comparison cases lack a well-defined interpretation.
Causal forests have great potential to quantify treatment effects in the presence of heterogeneity\textsuperscript{1}. A causal forest averages estimates of individual causal trees which can vary substantially and which makes it hard for an analyst to interpret how results for an individual data point are calculated and why this is meaningful. In this application paper, we are interested in understanding the ecological impact of World Bank aid projects. We employ causal forests to compute heterogenous treatment effects for World Bank aid projects on vegetation, specifically the NDVI index. In order to explain the computed result for an individual project, we fit an explanation model to locally match with the causal forest results for the vicinity of a specific project of interest. We use a weighted linear regression model to explain the calculated causal effect for any particular project because regression models are commonly applied by subject matter experts in this application domain and their interpretation is well understood.

\textsuperscript{1}This chapter is in the form of a paper we submitted for reviewing and publication in 2018.
5.1 Introduction

Machine learning techniques have a track record to help analyze complex real world data sets. In this paper, we analyze a data set for World Bank aid projects of the last 30 years with respect to the environmental impact these projects had. The World Bank has funded thousands of projects in third world countries over decades and one of its main objective is to improve the living conditions of humans in selected areas of the world. Often, economic development also impacts ecology. We are interested in evaluating the ecological impact of World Bank projects as measured by satellite data for an index of vegetation (NDVI). In [63],[64], the impact of World Bank projects is analyzed with classic econometric methods. In this paper, we apply causal inference techniques. This work is based on the Rubin Causal Model [41] (or potential outcome framework), where causal effects are estimated through comparisons between observed outcomes and the “counterfactual” outcomes one would have observed under the absence of an aid project [5]. Causal treatment effects in the presence of heterogeneity can be computed with causal forests. While the overall approach resembles the calculation of random forests, it differs in a number of important details, most prominently in the splitting rules for nodes, i.e., which covariate to select, which numerical threshold to use for the selected covariate’s value, and when to stop splitting a node. It shares a number of parameters with random forests, such as the number of trees, but also has additional ones as the minimum number of treated and untreated cases required per leaf node. While a causal forest excels at calculating heterogeneous causal effects, its black box nature makes the interpretation of results difficult for an analyst. However, the credibility of results goes hand in hand with the level of insight such results can provide. For example, [25] documents that a decision maker seeks explanations beyond predicted numbers. As we apply causal forests to our World Bank data set, several key questions naturally arise. Can we trust the calculated numerical values for causal treatment effects? What are the driving
forces for a treatment effect to be high or low?

While machine learning techniques have a track record of delivering amazing results, communicating how specific results are calculated or providing additional insights towards validation and verification of results is still subject to ongoing research.

We can check variable importance to the outcome to explain the results, [44], [78],[71] proposed different solutions.

In [26],[2],[4], we can see explain the model can help increase AI safety. For example, in [42], [2], they use influence function for explanation, in [45], the author give a comprehensive discussion of interpretability in machine learning research. Baehrens et al [6] propose an explanation for classifier. In [47],[58], the authors proposed generic frameworks to explain results produced with machine learning techniques. Ribeiro et al [58] describe a general framework to calculate simple explanation models that are open to human interpretation as a local approximation for a more powerful, global model that is difficult to interpret. We follow their guidance and evaluate ways to fit a weighted linear regression model to explain a causal forest model derived for our World Bank data set. Fundamental to this approach is the definition of a distance function for a pair of data points.

The contributions of this paper are: (1) we identify a spectrum of possible distance functions for causal forests and describe several examples of distance functions on this spectrum, different sampling methods for data points used in the fitting procedure for the local approximation, (2) we extend the LIME framework by taking causal inference into consideration, (3) we combine these features into an overall procedure to deliver a linear regression model to explain the calculated results of a causal forest, and (4) we apply this to the World Bank data set and discuss the resulting explanations for two concrete example project, namely one treated and one control case.

The paper is structured as follows. In Section 2, we briefly recall the concept of causality,
the calculation of a causal tree and causal forest. We also introduce different notions of distance that are essential for the computation of a local approximation with a model that is simpler to explain. We recall the concepts of linear regression as the explanation model we consider. In Section 3, we apply the approach to the World Bank data set and report our findings. We use two specific projects to highlight our analysis results. We conclude in Section 4.

5.2 Methods

In order to recognize causality in our data, we follow Rubin’s potential outcomes framework and consider a causal tree as in [5], [82]. Following Athey and Imbens’ notation, we consider a finite set of data points, each data point $i$ has a feature vector $X_i \in \mathbb{R}^n$ and a specific outcome $Y \in \mathbb{R}$. A data point $i$ may by subject to treatment, a case where we denote the outcome as $Y_i^{(1)}$, or otherwise it is $Y_i^{(0)}$. This distinction allows us to define the treatment effect $\tau$ at $x$ as $\tau(x) = E[Y_i^{(1)} - Y_i^{(0)} | X_i = x]$. The definition leads to the well-known dilemma that we can only observe $Y_i^{(1)}$ or $Y_i^{(0)}$ for each real entity in our data but not both as the entity either receives treatment or not.

5.2.1 Causal Tree and Causal Forest

To calculate an estimate for $\tau(x)$, Athey and Imbens derive a causal tree much like a regression tree but with a different splitting rule and different calculation of the estimated outcome, here $\hat{\tau}(x)$. We briefly recall the calculation of $\tau(x)$, which uses an estimated propensity $\hat{e}(X_i)$ for the probability to receive treatment, i.e. $e(X_i) = Pr(W_i = 1|X_i = x)$. Each node $s$ of the tree has an associated set of data points $L_s = L_s^{(0)} \cup L_s^{(1)}$ where some entries have been treated $L_s^{(1)}$, others not $L_s^{(0)}$. One can estimate $\hat{\tau}_s(x)$ as the difference
between sets $L_s^{(0)}$ and $L_s^{(1)}$:

$$\hat{\tau}_s(x) = \frac{1}{c_1} \sum_{X_i \in L_s^{(1)}} Y_i^{(1)}/\hat{e}(X_i) - \frac{1}{c_0} \sum_{X_i \in L_s^{(0)}} Y_i^{(0)}/(1 - \hat{e}(X_i))$$

where $c_0 = \sum_{X_i \in L_s^{(0)}} (1 - \hat{e}(X_i))$ and $c_1 = \sum_{X_i \in L_s^{(1)}} \hat{e}(X_i)$ are normalizing constants. Figure 5.1 illustrates set $L_s^{(0)}$ of control units and $L_s^{(1)}$ with treated units for a leaf node $s$ in a causal tree. The tree generation algorithm uses a splitting rule that looks for the covariate $x_i$ and threshold constant $c_t \in \mathbb{R}$ value for a condition ($x_i < c_t$) that maximizes the difference between $\hat{\tau}_s(x)$ for the parent node $s$ and its children $s'$ and $s''$, i.e., $\text{argmax}_{x_i, c_t} (\hat{\tau}_s(x) - (\omega_{s'} \hat{\tau}_{s'}(x) + \omega_{s''} \hat{\tau}_{s''}(x)))$. $\omega_{s'}$ is the percentage of data points that go into child node $s'$, $\omega_{s''} = 1 - \omega_{s'}$ is the corresponding weight for $s''$. As an additional side condition, the splitting rule does only apply if resulting nodes $s'$ and $s''$ each hold at least $k$ data points of treated and untreated cases. We choose $k = 10$ in our analysis of the World Bank data set.

We randomly selecting 80% of the covariates and 80% of the data to grow each causal tree in the causal forest. As we randomly draw the subset of covariates and data points independently for each tree, we follow a common approach to build a random forest to obtain a causal forest in our context. For each data point $i$, we calculate $\hat{\tau}(x)$ by identifying the leaf node $s$, where the data point resides in, in each individual tree to calculate an estimate $\hat{\tau}_s(x)$ for that tree and then average $\hat{\tau}_s(x)$ values across all trees in the forest. While the calculation of $\hat{\tau}(x)$ is straightforward, the causal forest does not provide an immediate interpretation of it. For example, it is hard to explain what the most relevant covariates are for a particular estimate of $\hat{\tau}(x)$. For further details on causal trees, forests and their implementation within the scikit-learn framework, see [5], [82], [91], [90].
5.2.2 Computation of an Explanation Model

In [58], Ribeiro et al propose the LIME approach to use a simple model to explain a more complex model via local approximation with data points close to some point of interest $x$. Formally, LIME solves

$$\xi(x) = \text{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where $f$ is a function expressed with a complex model, in our case $f = \hat{\tau}(x)$ and is computed with a causal forest, $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a model interpretable and much simpler than $f$ that approximates $f$ in the vicinity of a reference point $x$ and $\pi_x : \mathbb{R}^n \rightarrow \mathbb{R}$ describes a weight function that decreases with an increase of the distance between any given points $z$ and $x$. Finally $\Omega(g)$ denotes the complexity of $g$. So, in plain English, LIME looks for the least complex $g$ that also has a small value for loss function $L$. We can find this by considering a weighted square loss function $L$ that uses sample points $z$ to quantify how well $g$ approximates $f$.

$$\eta(f, g, \pi_x) = \sum_z \pi_x(z)(f(z) - g(z))^2$$

(5.1)

In this equation, let $\pi_x(z) = \exp(-d(x, z)^2/\sigma^2)$ be an exponential kernel defined with the help of a distance function $d(x, z)$ and width $\sigma$.

In our context, this means that we can select a particular data point $i$ - a World Bank project - and then calculate a model, here a linear regression model, that is fitted to the local vicinity of $X_i$. LIME requires us to resolve two issues, namely to decide how to sample data for the computation of the local model and also to define a notion of distance $d(x, z)$ between feature vectors of data points. We discuss options for a distance calculation first. For simplicity of notation, we will simply refer to a data point as $x$ if we consider a data
point with feature vector $X$.

5.2.3 Distance Measurement

Fitting a simple model to locally approximate another, more complicated model as proposed in [58] relies on a notion of distance between two data points. This is formalized by a distance function $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ that maps feature vectors of two data points in $\mathbb{R}^n$ to a non-negative value. In the following, we discuss different types of distance functions, namely one that is based on the structure of the complex model (here: causal forest), one that obtains guidance from the structure of the complex model to weigh differences in value across covariates and one that is purely based on difference in value across covariates.

5.2.3.1 A Purely Structural Notion of Distance

A distance function can be purely based on the structure of the complex model. In our case, we consider a causal forest that is built from individual causal trees. For a single tree $t$ with a set of nodes $S$, let function $l_t : S \rightarrow \mathbb{N}$ describe the level of node $s$ in tree $t$, i.e., $l_t(s) = 0$ if $s$ is the root node and $l_t(s) = l_t(s') + 1$ if $s$ is not the root node and node $s'$ is the parent node of $s$ in tree $t$. Each node $s$ has a path from the root node to $s$. For any two nodes $s$ and $s'$ in a tree, their corresponding paths share a common prefix that starts with the root node. Let their longest common prefix end at some node $r(s, s')$, which may or may not be the root node itself.

Let $s_x$ denote the leaf node in the tree whose bin contains data point $x$, similarly for $s_y$ and data point $y$. So, node $r(s_x, s_y)$ is the root node of the smallest possible subtree that contains $s_x$ and $s_y$. We can interpret a tree as a graph and count the number of steps or edges between nodes to define a distance, i.e., $d_t : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{N}_0$ with $d_t(x, y) = l_t(s_x) + l_t(s_y) - 2l_t(r(s_x, s_y))$. 

91
As all data points reside in leaf nodes, the corresponding paths to these nodes can share a number of edges which correspond to properties for the covariates that are checked on these tree nodes. If $x$ and $y$ are in the same leaf node, all conditions on covariates that are expressed along the path of the tree from root node to the leaf node are fulfilled by $x$ and $y$ in the same manner, which makes them similar in this respect and evaluates for $d_t(x, y) = 0$. The fewer conditions $x$ and $y$ share, the larger is $d_t(x, y)$ with $l_t(s_x) + l_t(s_y)$ as an upper bound for their distance.

We can extend the distance of $x$ and $y$ from a single tree $t$ to a forest $T$ by any measure that aggregates the information of a set of distance values, typical candidates are minimum, maximum, average or median. For the average distance, we define $d_T(x, y) = \sum_{t \in T} d_t(x, y) / |T|$ which matches with the calculation of the causal treatment effect from forest $T$.

### 5.2.3.2 A Purely Data-driven Notion of Distance

If we want to include the particular value settings for each data point, we can take advantage of the fact that each data point is a vector of real-valued covariates. We can use any norm on the difference $x - y$ to measure a distance, for instance, a p-norm with $p=1$ for the taxicab norm, $p=2$ for Euclidean distance, or $p = \infty$ for the maximum norm. We focus on the Euclidean distance for illustrating purposes, so $d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$. This will introduce common problems such as a need for normalization if numerical values across covariates differ in scale and one considers numerical differences across all covariates of equal importance. This leads to the question if some covariates should matter more than others, which can be formalized with a weighted summation $d(x, y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$ with $w_i \in [0, 1]$.

So as a general option, one can let the analyst decide on covariates that matter for the
distance calculation by assigning weights \( w_i \) for each \( i \). For example, in our application, the data naturally has a geographical location. Covariates include longitude and latitude, such that \( w_i = 1 \) for longitude and latitude and 0 for other covariates resembles the notion of spatial proximity.

5.2.3.3 A Combined Notion of Distance

The two classes of distances discussed so far are extreme cases on a range of possibilities: the first distance only relies on structural information contained in the complex model, the second only relies on properties of the considered data set. If we want to combine structural information with the numerical values of covariates, we can obtain weight settings in various ways.

We can select weights according to paths in trees. Let \( P(s) = \{ s_0, s_1, \ldots, s_c \} \) where \( s = s_c \) be the path of nodes from the root node \( s_0 \) to a node \( s_c \) in a tree. So for \( d_l(x, y) \) with a weighted summation, we can define weights \( w_i = 1 \) if \( x_i \), resp. \( y_i \) corresponds a node in \( P(s_x) \cup P(s_y) \) and 0 otherwise. This definition of \( d_l(x, y) \) would account only for covariates that are on any of the two paths to \( s_x \) or \( s_y \). This definition has a lot of room for variation. A direct variation would consider the intersection \( P(s_x) \cap P(s_y) \) instead of \( P(s_x) \cup P(s_y) \). Of course, one can select other values for weights \( w_i \), e.g., weights that decrease with the level \( l_i(s_i) \) of the node in the tree as nodes closer to the root node may represent covariates that matter more for the calculation of causal effect. Alternatively, we can calculate average weights \( \bar{w}_i \) across forest \( T \) from weights \( w_i \) calculated per individual tree for \( x \) and \( y \) and then in turn use \( \bar{w}_i \) for the calculation of \( d_l(x, y) \).

Finally, we can derive further information from causal forests that can be used for weights. The splitting rule for a causal tree is based on a measure of purity such that each node in a tree \( t \) has an associated purity value (for instance information gain or Gini
impurity). For a single tree, we can calculate purity values for each covariate (summation of values if several nodes correspond to a covariate). For the whole forest, we can calculate an average purity value for each covariate and then use that value as a weight for the covariate in the distance measure.

5.2.4 Sampling Data for an Explanation Model

We assume that we are given a causal forest, a specific data point \( x \) of interest, and a target number of data points that the computed sample neighbor set \( NH \) should provide. As LIME weighs samples by their distance from \( x \) and gives higher weight to samples close by, LIME does not put a hard constraint on a sampling procedure for proximity. Consequently, a simple sampling procedure which we call tree sampling can just randomly select other data points \( y \) to obtain set \( NH \). The causal forest is then used to compute for each \( y \in NH \) the average distance \( d(x, y) = 1/|T| \sum_{t \in T} d_t(x, y) \) over all trees in the forest where \( d_t(x, y) \) denotes the distance between \( x \) and \( y \) in tree \( t \).

As tree sampling does not take into account that LIME has a preferences for samples with small distance to \( x \), one can obviously vary the sampling method to do so. This depends on the definition of distance \( d_t(x, y) \). For cases where the structure of the tree \( t \) matters, data points \( x \) and \( y \) that reside in nodes \( s_x \) and \( s_y \) which share a path from the root node are typically closer to each other. In the extreme case where \( x \) and \( y \) are in the same leaf node \( s_x = s_y \) one can expect the distance to be small. So for leaf sampling, we look for data points that share the same leaf node with \( x \). Across all trees, we count how often each data point is sharing the same leaf node with \( x \) and thus, after a normalization, we obtain an empirical distribution in this way. We then sample the desired number of data points from this empirical distribution. Note that the support of the empirical distribution gives an upper limit for the cardinality of \( NH \). As for tree sampling, we calculate the
average distance $d(x, y)$ between $x$ and each selected data point $y$. Note that $x$ and $y$ are in the same leaf node for at least one tree in the forest, but not necessarily for all trees in the forest.

Alternatively, one can perform leaf sampling by randomly selecting a fixed number of data points from leaf $s_x$ for each tree in the forest, i.e., each tree contributes a set of candidates that are sampled from $s_x$. We use the latter approach in our analysis and sample one candidate per tree from a forest $T$ such that $m \leq |NH| \leq |T|$ where $m$ is the minimum leaf size used in the splitting rule. Since several trees can propose the same candidates $|NH| \leq |T|$.

There are several parameters we need to tune for this calculation, such as the cardinality of $NH$, but also in the causal forest calculation, the minimum cardinality of treated and untreated sets per leaf node in each causal tree.

**Figure 5.1**: Distance in a causal tree
5.2.5 Sparse Linear Explanations

A generic linear explanation has the form

\[ h(x) = \beta_0 + \sum_{i \in F} \beta_i x_i \]  (5.2)

and the coefficients \( \beta_i \in \mathbb{R} \) inform us about the impact of one unit change of \( x_i \) on the outcome \( h(x) \). In order to use a linear explanation in the context of Rubin’s model, we recall its fundamental equation

\[ \tau(x) = Y(x)^{(1)} - Y(x)^{(0)} \]  (5.3)

where \( x \) characterizes the particular project and the superscript indicates treatment. In the above equation, \( Y(x)^{(1)} \) and \( Y(x)^{(0)} \) are to be estimated with a linear explanation model. To do so, we include covariates from a feature set \( F \) that is a subset of all covariates and an indicator variable \( x_T \). The indicator variable \( x_T = 1 \) for treated and 0 for control. This leads to

\[ h(x) = \beta_0 + \sum_{i \in F} \beta_i x_i + \beta_t x_t \]  (5.4)

such that

\[ \tau(x) = Y(x)^{(1)} - Y(x)^{(0)} = h(x|_{x_t=1}) - h(x|_{x_t=0}) = \beta_t \]  (5.5)

So \( \beta_t \) in (5.4) is the causal effect and we can now interpret (5.4) following common lines of argumentation to interpret a linear regression model to better understand the causal effect computed by a causal forest.

Since \( h(x) \) is expected to be a good approximation only close to the reference data point, we can use \( h(x) \) to explain how small changes to a covariate setting would impact the causal effect, i.e., we can interpret the sign of \( \beta_i \) and its magnitude. By comparing coefficients across covariates, we can recognize if there is a subset of dominating covariates that are
mainly responsible for the outcome and which covariates are in that set. Similarly, we can identify covariates whose absolute value is close to zero and thus have a rather negligible impact on the outcome. For linear regression models, it is also common to test if coefficient values are statistically significant and we can do this here as well.

In the context of LIME, \( g(x) = h(x|z_t=1) - h(x|z_t=0) \) and we can compute \( h(x) \) with help of a weighted linear regression. We need to decide on the data set \( NH \), the selection of weights, and the set of covariates \( F \). To summarize how the explanation model is derived from the causal forest for a particular project of interest \( x \):

1. The set of trees \( T \) of a CF provides a distribution of values for \( \hat{\tau}(x) \), which creates an opportunity for outlier detection and removal and possible reduction of \( T \).

2. For \( T \), we perform leaf sampling to obtain a set of neighbors \( NH \) of treated and control cases for \( x \).

3. The distance function \( d(x,y) \) is a combined notion of distance with average weights obtained for covariates on paths \( P(s_x) \cup P(s_y) \) across all \( t \in T \) and \( y \in NH \).

4. An exponential kernel is used to transform \( d(x,y) \) into a weight function used in the weighted linear regression.

5. The CF provides us with an estimate for \( \hat{\tau}(x) \) that equals \( \beta_t \) in the linear regression model.

The CF computation selects certain covariates and thresholds to bin projects that are then used to estimate \( \hat{\tau}(x) \) for some project \( x \). The model transformation we chose carries much of this information over to an explanation model that is a local approximation of CF with respect to \( x \) and that is open to the common interpretation of a linear regression model.

With this information, we can explore at least three avenues to work with the explanation model. Firstly, we can solve (5.4) with the given data points, weights, and fixed
\( \beta_t = \hat{\tau}(x) \) for \( \beta_0, \beta_1 \ldots, \beta_{|F|} \) as a weighted linear regression. The resulting function \( h(x) \) is a representation of the CF with respect to the project of interest \( x \).

Secondly, we can solve (5.4) as for the first option but consider \( \beta_t \) as a coefficient that is computed with the weighted linear regressions just like \( \beta_0, \beta_1 \ldots, \beta_{|F|} \). If \( \beta_t \) and \( \hat{\tau}(x) \) differ significantly, we recognize that some information between CF and the explanation model is missing. One prime suspect is that some covariates interact, which is a common limitation of the basic linear model that we use for (5.4). This leads to the third direction to proceed.

Thirdly, we can iteratively add interaction terms \( \beta_{ij}x_ix_j \) to (5.4), one term at a time, to explore if this enhanced model yields a better fit of \( \beta_t \) and \( \hat{\tau}(x) \). This leads to a search procedure (an optimization problem) to find a linear equation model such that \( |\beta_t - \hat{\tau}(x)| < \epsilon \) with the smallest number of interaction terms.

Note that the point of the exercise is not to find the ground truth for the data set but to compute an explanation model that allows us to interpret and understand the computed result of the causal forest.

### 5.3 Causal Effect Calculation for World Bank Data

The World Bank data set is an enhanced data set that adds geo-location data to aid projects, a time series of vegetation data based on satellite imagery (NDVI: Normalized Difference Vegetation Index), a population estimate based on night light satellite data, as well as weather data in terms of time series data for temperatures and precipitation. The aid project data provides project duration and funding levels. We consider individual project locations as data points such that we have \( N = 16369 \), which we will consider projects for simplicity. In the data set, we consider a high amount of funding a treatment in the sense of the Rubin causal model, such that the 33% of projects with highest amount of funding form the treatment group and the rest the control. For treated cases, the beginning year
of a project is used to recognize the end of the time series data that is not influenced by
the treatment and can be used as a covariate. To measure the outcome, we look at the
difference between average NDVI values before and after the beginning of a project. We
reduce time series data with the help of a linear regression to covariates for intercept and
slope. This helps us to reduce the number of covariates to consider to 36.

We perform a calculation of causal effects \( \hat{\tau}(x) \) with the help of a causal forest. We use
a configuration with 500 trees, where randomly chosen 80% of the data are made available
during tree generation and 40% of features for each node split. The overall distribution
of \( \hat{\tau}(x) \) values shows that projects have a slightly negative impact on vegetation. This
is plausible as the typical objective is an improvement in human living conditions and
economic development and in this sense having an almost marginal negative impact is an
acceptable and welcome result. The causal forest as a black box model is not open to
some direct interpretation, in particular, if we seek some explanations for the result of a
particular project location.

We consider two specific projects, the Zhejiang Urban Environment Project (ZUEP)
and the Municipal Infrastructure Project in Lebanon (MIPL).

5.3.1 Zhejiang Urban Environment Project

According to the World Bank’s project description, the objective of the Zhejiang Urban
Environment Project (ZUEP) is to “enhance the efficiency and equity of waste management
in Ningbo and Hangzhou municipalities and the redevelopment of the historic inner city in
Shaoxing and Cicheng, and thereby facilitating the sustainable development of these cities
and establishing a model for conservation of cultural heritage in other Chinese cities” [8].
The causal forest provides us with an estimate \( \hat{\tau}(x) = 0.0002 \) for the causal effect but no
further explanation.
As a first step to better understand the result, we recognize that each tree of the causal forest contributes a single estimate value and that $\hat{\tau}(x)$ is the average of these values. So, we can look into the distribution of values contributed by the set of trees, which we show in Fig. 5.2 for ZUEP. The histogram in Fig. 5.2 indicates a unimodal distribution with little dispersion, which lets us conclude that the vast majority of trees are in good agreement on the resulting value, or in other terms that the data set does not have so much variability that the random selection of covariates and data points leads to vastly different trees. A bimodal or multimodal distribution would indicate that subsets in the data set exist that on their own would give very different estimates for $\hat{\tau}(x)$ and the estimate is sensitive to the selected data.

From a practical point of view and for the case of ZUEP, we see that we can remove the outliers in the histogram in Fig. 5.2, which in turn implies to remove the subset of trees that produces these outliers and thus work with a reduced causal forest.

![Figure 5.2: Histogram of values contributed by trees in the CF for ZUEP to compute $\hat{\tau}(x)$](image)

The calculation of $\hat{\tau}(x)$ is based on the difference between changes in NDVI values
for treated and control projects. We use leaf sampling to obtain a set of sample neighbor projects $NH$ that are used for the calculation of the causal effect for this project and further analyze this set, e.g., we can visualize the projects’ locations as on the world map in Fig. 5.3. In this map, the ZUEP location is shown as a blue dot, other projects in NH are represented with red dots. For the weighted linear regression, each data point carries a weight that is visualized with the area of a dot in Fig. 5.3. The figure shows that spatial proximity matters as a majority of projects are reasonably close to ZUEP but spatial proximity can not be the only criteria as there are cases with huge differences in longitude and latitude that carry a significant weight for the linear regression, i.e., they are considered close to the project of interest. Obviously the criteria used by the causal forest to bin projects in leaf nodes is not simply spatial proximity.

![Figure 5.3: Geographic locations of a sampled subset of projects used in the estimation of $\hat{\tau}(x)$ for ZUEP](image)

This leads to the question which covariates do matter most for the calculation of $\hat{\tau}(x)$. The causal forest calculates results from a data set with 36 covariates and each single tree
selects particular ones on the path $P(s_x)$ to the leaf node $s_x$ that contains ZUEP. The average path length is five and much less than 36, so which covariates are selected for the calculation of $\hat{\tau}(x)$ for ZUEP?

In order to represent the selection of covariates in the forest, we count how often each covariate occurs on path $P(s_x)$ for a single tree and then sum this over all trees in CF. Figure 5.4 shows the corresponding percentages for project ZUEP when we group features according to what they measure, e.g. geolocation subsumes longitude and latitude. The grouping is for visualization purposes only. The 36 covariates are grouped as follows, geolocation subsumes longitude and latitude, continent, slope, elevation, and distance to rivers. Precipitation includes slope and intercept of the linear regression on annual averages, minimum and maximum for precipitation and temperature does this for the corresponding time series for temperature. Population absorbs covariates for the population in the year 1990, 1995, and 2000. Urbanization includes the average and trend of a time series on the estimated population density before the project started, the accessibility to urban areas, and the distance to the next major road and the next commercial river. Sun light accounts for the slope and intercept of the linear regression on a time series on day light before project the project started, NDVI for slope and intercept of the linear regression on annual average NDVI values before project started. The starting year is just a single covariate for the starting year of the project.

We can see that the geographical location, the natural environment (precipitation, temperature, sunlight) and the preexisting vegetation have a huge influence on the calculation of the causal effect. Human civilization aspects are less relevant. The CF is particularly suited to account for heterogeneity and the selection of these covariates indicates that these covariates provide a relevant context for projects to be comparable with ZUEP for the causal effect calculation.
Function $\hat{g}(x) = g(x) = h(x|x_t = 1) - h(x|x_t = 0) = \beta_t$. Function $h(x)$ is the outcome of the regression; it predicts the change in NDVI values $h(x)$ as a function of covariates and one additional variable $x_t$ that allows the distinction between treated cases ($x_t = 1$) and control ($x_t = 0$). Weights used in the regression are derived from the average distance $d(x, y)$ between ZUEP $x$ and $y \in NH$. We transform the distance to weights with the help of an exponential kernel such that a huge distance results in a weight close to zero and a small distance to a weight that is close to one. The outcome is shown in Fig. 5.5. So, it is the distance function that carries insights in which projects are close and thus good for comparison into the linear regression model. The subset of projects $NH$ considered for the regression is a second aspect that incorporates characteristics of the CF into the regression model, and finally, the subset of covariates considered in the regression is chosen according to the subset derived from the CF.

Function $h(x)$ is the explanation model and we can follow common approaches to in-
interpret the results. We can evaluate the impact of particular covariates based on sign and magnitude of corresponding $\beta$ coefficients. As a first step, we calculate a linear regression with fixed $\beta_t = \hat{\tau}(x) = 0.0002$ for ZUEP. From the outcome, we see for instance that the three coefficients that are statistically significant and are of largest magnitude are all temperature related, i.e., these are the covariates \textit{average temperature before project started} with $\beta_i = 0.731$, the \textit{intercept of a linear regression on average temperature} with $\beta_i = -0.63$, and the \textit{intercept of a linear regression on the max temperature} with $\beta_i = -0.54$.

As a second step, we calculate a linear regression with $\beta_t$ determined in the regression. In this case, $\beta_t$ shows us to what extend the linear regression model is consistent with the calculated results of the CF for ZUEP. We obtain values for $\beta_t$ as shown in Table 5.1 in row ZUEP. In Table 5.1, we see that the coefficient of the treatment feature $\beta_t$ is positive, indicating that the mean average NDVI values we can expect to see after the project started increases as the value of the treatment variable increase from 0 to 1. Given a test of the
null hypothesis, $H_0 : \beta_t = 0$ against the alternative $H_A : \beta_t \neq 0$, assuming that all the other features do not change, we get a test statistic $t = 3.623$, and the corresponding p value is close to 0 and much less than 0.05, therefore, we reject the null hypothesis at the 0.05 level of significance. Within the confidence interval [0.025,0.975], the coefficient ranges from 0.002 to 0.005. In conclusion, the coefficient is statistical significant and has a positive effect. However, the $\hat{\tau}(x)$ value is outside of the confidence interval values for $\beta_t$ such that we investigate if we can identify some interactions among variables that derive a better explanation model. From a first iteration that goes through the $n \times (n - 1)/2$ combinations of adding a single interaction term $\beta_{ijx_ix_j}$, we obtain the best fit for covariates intercept of a linear regression on annual average NDVI values and intercept of a linear regression on average light values, with a coefficient 0.0018 that is statistically significant. The remaining difference between $\beta_t$ and $\hat{\tau}(x)$ is reduced to 0.0016. Obviously one can push this further to reduce this difference even more by adding more interaction terms for the price of a more complex linear regression model. As in [76], we can find such interacted covariates.

The lesson we learned when applying the weighted linear regression model as the explanation model is that the calculation is sensitive to the selection of weights. Our choice of an exponential kernel follows the literature [58] but it is not the only possibility. We leave it to future work to explore further options for other kernel functions and their most suitable parameter settings. For example, many weights in Figure 5.5 are close to zero, which suggests that its corresponding data points could most likely be omitted.

On a separate note, the tree generation for CF is straightforward to parallelize, which

Table 5.1: WLS Regression Results of Treatment feature for projects ZUEP and MIPL

<table>
<thead>
<tr>
<th>Project</th>
<th>$\beta_t$</th>
<th>t test</th>
<th>p value</th>
<th>[0.025</th>
<th>0.975]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZUEP</td>
<td>0.0034</td>
<td>3.623</td>
<td>0.000</td>
<td>0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>MIPL</td>
<td>-0.0160</td>
<td>-3.081</td>
<td>0.002</td>
<td>-0.026</td>
<td>-0.006</td>
</tr>
</tbody>
</table>
helped us to significantly speed up calculations for this experiment.

5.3.2 A Municipal Infrastructure Project in Lebanon

For a second example, we select one that received little funding and is thus considered a control and not a treated case in our data set. The CF allows us not just to evaluate treated cases but to also obtain causal effect estimates for controls and derive an explanation model accordingly.

According to [7], the objective of the First Municipal Infrastructure Project in Lebanon (MIPL) is to improve access to basic services at the local level, which "has been achieved through building and/or upgrading of essential infrastructure; construction of about 3,350 km of roads; 305.7 km of retaining walls; installation of 19,767 streetlight poles; improvement of 290 km of storm drainage networks; improvement of 28 km of potable water networks; and rehabilitation of 36 km of sewerage networks. 17 municipal building were reconstructed, in addition to the construction of 15 public facilities in 15 municipalities." Improving infrastructure facilitates economic activity, which can have different effects on vegetation. For an area with rich vegetation, the change in NDVI is most likely negative as the natural vegetation may be replaced with buildings in urban areas and monoculture farm fields in rural areas. For a semi-arid region, infrastructure may increase farming that relies on irrigation to grow agricultural crops and in turn improve the NDVI value for that area.

The causal forest estimates the causal effect for this project as \( \hat{\tau}(x) = -0.0003 \), which is the mean value of the set of values obtained from all trees in the forest that are shown as a histogram in Fig. 5.6. The distribution is similar to the one for ZUEP in Fig. 5.2: a unimodal distribution with little dispersion and a few outliers that one can consider for removal.
Figure 5.6: Histogram of values contributed by trees in the CF for MIPL to compute $\hat{\tau}(x)$

Visualizing the spatial distribution of projects in NH on a world map in Fig. 5.7 shows that a vast majority of projects are close by and very few projects of substantial weight are further away. Compared with Figure 5.3, the set of neighbors is rather dense and the MIPL project is in the center of a subset of substantial size.

In Figure 5.8, we observe similar results as for ZUEP before. The only significant difference apparently is that NDVI lost about 11%, the geolocation group of covariates gained more than 5% and some minor increases towards population and urbanization related covariates. Transforming the distance values into weights, we see a distribution of values as shown in the histogram in Fig. 5.9 with many data points considered too distant to be of significant weight.

Similar to ZUEP, we can calculate a linear regression for $\beta_t = \hat{\tau}(x)$ to obtain an explanation model for the CF result. We can then follow up with a calculation of a linear regression where $\beta_t$ is determined in the calculation to check the consistency with $\hat{\tau}(x)$ obtained from the CF. Row MIPL in Table 5.1 shows that the treatment coefficient of $-0.016$ is negative
indicating that the average NDVI values after a project launch would decrease as the value of treatment variable changes from 0 to 1. We get a test statistic of $t=3.623$ and p value of 0.002, which is less than 0.05. Therefore, we reject the null hypothesis at the 0.05 level of significance. Within the confidence interval $[0.025, 0.975]$, the coefficient ranges from -0.026 to -0.006. Similar to ZUEP, the value of $\hat{\tau}(x)$ is not included in the interval and we would one more time go through possible interaction terms to reduce the difference between $\beta_t$ and $\hat{\tau}(x)$.

In this control case, the causal effect is negative, which suggests that if more funding is invested in this project, the project outcome would reduce the vegetation index.

### 5.4 Causal Effect of World Bank Projects to Night Lights

The aim of World Bank projects is to improve people’s living standard and reduce poverty. As the night lights is a good indicator for the level of poverty. The area with larger night
Figure 5.8: MIPL feature frequency in CF

Figure 5.9: Distances scaled by kernel function for MIPL

lights values are thought to be more rich than the area with smaller night light values. Therefore, besides the research of World Bank projects to the change of environment, we
can also analyze the impact of World Bank projects to the change of poverty, which is also the goal of projects.

![Figure 5.10](image.png)

**Figure 5.10**: Distribution of change of nights lights

In figure 5.10, we can observe that the night light changes for most of the projects are positive and in figure 5.11 shows that most of the projects have positive impact to the change of night lights, which indicate that World Bank projects do have positive impact to reduce poverty.

### 5.4.1 Discussion

One of the key insights is that the explanation model is only required to be an accurate and interpretable representation of the CF results for a particular project and not of the ground truth, i.e., if the CF estimation is poor, the explanation model should be consistent and thus also be a poor estimation of the true causal effect values.

We recognized that the weights in the weighted linear regression influence the outcome a lot, so the way covariates are identified and rescaled with a kernel impacts the overall
outcome. As there are multiple kernels possible, further insights are needed on which one to select in which situation.

The multicollinearity in data may render a linear regression without interaction terms very inaccurate, recognizing the need for more interaction terms by the difference between $\beta_t$ and $\hat{\tau}(x)$ and exploring the benefits of interaction terms is possible in this context.

Trimming data by outlier removal is possible in various stages of the procedure, i.e., in the selection of trees from the CF to consider, in the selection of elements for NH based on their weights. This definitely simplifies and accelerates the computation of results but further analysis is needed to identify meaningful thresholds for removal.

Given that we obtain a distance function from the CF, one can consider a selection of $k$ nearest neighbors for NH instead of sampling data points from leaf nodes $s_x$ in each tree.

A CF naturally describes step functions while the linear regression derives a continuous function. This difference may impact the accuracy of the overall approach.

The fact that the calculation of each tree for the CF is independent from the calculation
of another tree, it is natural to run the tree generation in parallel in order to speed up its calculation.

5.5 Conclusion

In this paper, we are interested in quantifying the contribution of World Bank Aid projects to changes in the natural environment. We employ causal forests to calculate the causal effect that aid projects have to changes in vegetation. As with many machine learning techniques, the causal forest results are difficult to interpret beyond their plain values. To gain a better understanding what drives the causal forest calculation we compute an explanation model following the LIME approach that provides insights into causal effects for individual World Bank Aid projects. We choose to use a linear regression model as subject matter experts are experienced in using and interpreting regression models in this particular domain. We focus on two particular projects, the Zhejiang Urban Environment project in China and the Municipal Infrastructure project in Lebanon. The explanation models indicate that the projects themselves have a mild impact on the environment and that the causal forest calculation identified preexisting conditions for climate and vegetation as driving factors for change.
Chapter 6

Conclusion

In this thesis, we conducted a causal inference analysis for the World Bank dataset and investigated World Bank projects' heterogeneous impacts on a change of environment. First, we used the potential outcome framework to formulate the causal effect estimation problem for the given World Bank dataset. Second, we estimated the causal effect of each World Bank project on the change of environment using a refined causal forest. Finally, we proposed an explanation model to interpret the causal effect calculated from a causal forest for each World Bank project.

To enable estimation of the causal effect for each World Bank project with the potential outcome framework, we need both treated and control data in the dataset. We can assign a unit to the treated group if the unit represents a World Bank project, and assign a unit to the control group if the unit does not. However, we do not have such data that can be assigned in the control group given the World Bank dataset collected by Aiddata. Therefore, we used the covariate funding as a threshold to separate the dataset into treated and control groups, assigning projects with larger funding into the treated group and projects with smaller funding into the control group, under the assumption that larger funding projects would have bigger impacts on the change of environment. We also used a binary treatment
in this thesis. However, the treatment level can be more than two as described in [80] and [48], where they used multiple treatment levels to estimate the causal effect. We can use multiple treatment levels for the World Bank dataset as well, and we leave this for future work.

To calculate the causal effect for World Bank projects, we need outcome and confounders as well as treatment assignment. We used NDVI, a common metric for vegetation, to represent the environment. As projects have different starting years, the outcome used in this thesis is defined as the difference between the average annual NDVI values for before and after the projects started. For time series confounders (e.g., temperature, NDVI, precipitation), we had to transform them before we used them in a causal tree or a causal forest because projects with different starting years are not comparable without transformation. We fitted a linear regression model for the time series covariates before the starting years, and represented them with slopes and intercepts of the linear regression models. We replaced the times series in the dataset with the intercepts and slopes. Moreover, for each project, we only considered its time series covariates before the project starting year, because they may be affected by the treatment.

We relied on a causal forest to estimate the heterogeneous causal effects of the World Bank projects. Very little research has been done on the application of a causal forest to observational data. Typically, this research has focused on synthetic data rather than a real dataset. For example, the causal forest is investigated with synthetic data generated from a simple stochastic model based on normal distributions as described in [82]. We investigated the performance of a causal forest further with synthetic data, which was generated from a refined stochastic model that considered the influence of the number of covariates and distributions fitted to the World Bank dataset. We evaluated the impact of the number of confounders on the accuracy of a causal forest, and we observed that the errors in the
causal forest increased when more confounders are used in the model.

Leaf size, which is the minimum number of units in a leaf node, is a parameter used in the causal forest algorithm. The data in a leaf node are used to estimate the treatment effect. The optimal leaf size depends on the quality of data. For a quality dataset with highly comparable units, a causal tree with a small leaf size will achieve a better accuracy. But for an observational data set, a large leaf size is preferred such that averaging can alleviate the impact of individual data points. In practice, if the causal effects of a unit vary a lot in a causal forest, then this may indicate there are not enough comparable units in the dataset and the quality of the dataset may be low.

In many applications, decision making does not rely on statistical point estimates alone but considers their confidence intervals as well. In the literature, some research has been done on investigating confidence interval for forest-based methods. For example, in [82], the researchers proposed a method for estimating confidence interval for a causal forest. In [83], they studied the confidence interval of a random forest based on jackknife and the infinitesimal jackknife. However, confidence intervals are not studied for causal forests with observational data. Confidence intervals of the causal effect of a World Bank project estimated from a causal forest is not studied in this thesis, and we leave it for future work.

The causal forest is a black box model, and it is not easy to interpret its results. Even for a single causal tree, we cannot explain its results in a straightforward manner as for a decision tree. For a decision tree, we can interpret the results using the features and the thresholds along the path from root node to a leaf node. However, for a causal tree, the features and the thresholds are used for separating data into more homogeneous subspaces for control and treated subsets. To interpret the causal impacts estimated from a causal forest for any fixed project of interest, a linear regression model need to be learned based on the projects nearby that project in our case. In the linear regression model, only features
that appear most along the path from root to the project and its neighbors in the causal forest are used because the features contribute to the outcomes are different from project to project. Besides using the features alone, we can also assign weights to the features. The weights can be measured by the frequency of their appearance in the causal forest. Also, variable importance, which is often used in a random forest, can be an alternative method to weigh the features. In addition, the structure of the tree can be taken into consideration. Features shown close to the root node receive higher weights than the features appearing closer to a leaf node.

In the explanation model, since the treatment is included as a binary variable, we should be careful when using a regularization method such as Lasso and Ridge, because they may introduce bias to the coefficient of the treatment variable, which will lead to a wrong interpretation. Moreover, the results are sensitive to weights, as they are calculated by the transformation of distances by a kernel function. For different applications, kernel functions may vary depending on the distance metrics.

In this thesis, we hold the SUTVA assumption that the outcome of a unit will not be impacted by the treatment of other units. In other words, the treatment effects would not spill over between treatment and control group. In practice, a project may impact projects around it. For instance, for hospital building project, besides people living within the project area, a control group of people living in the nearby area can also benefit from the hospital. The causal effect of the hospital cannot be correctly estimated without taking a spillover effect into consideration because the outcome of the control group is impacted by the spillover treatment effect. In [66], the researchers investigated causal effects with the spillover effect taken into consideration, but they did not consider the heterogeneity of the causal effects. We leave it for future work to take both heterogeneity and spillover into consideration.
In conclusion, the contributions of this thesis include the following: 1) Because World Bank projects have various starting years, time series data cannot be used in a causal forest directly. We transformed these time series data with a linear regression model and represented the data with slope and intercept of the fitted linear regression model. 2) To measure the causal effect of a World Bank project, we used the covariate funding as a threshold to separate the data, assigning projects with larger funding into the treated group and projects with smaller funding into the control group, under the assumption that larger funding projects would have bigger impacts on the change of environment. 3) There is very little research on the application of the causal forest approach, and factors that influence the performance of a causal forest are not well studied. Therefore, we investigated the performance of a causal forest with synthetic data, which were generated using a fitted distribution of World Bank data. We investigate the impact of the number of confounders, leaf size, and data size to the causal forest. 4) We interpreted the causal effect of any fixed project of interest by its neighboring projects. We used dynamic distance metrics to find the neighbors of a project of interest, and learned a linear regression model using these neighbors. Then, we were able to interpret the linear regression model as it is common practice, because linear regression models are widely used and well understood by analysts.

Using the research presented in this thesis, developers and policy makers can estimate the causal effect, as well as the reasons for the causal effect, for a given unit (a World Bank project in our case). Knowing these causal effects and the reason behind will contribute to a more beneficial, effective project planning in the future.
Bibliography


[8] **World Bank.** Undertake critical reforms to improve Lebanon’s fiscal stability, create employment, provide adequate infrastructure, and support social inclusion, 2014.

[9] **Claudia N Berg, Uwe Deichmann, Yishen Liu, and Harris Selod.** *Transport policies and development (English).* Policy Research working paper; no. WPS 7366; Paper is funded by the Knowledge for Change Program (KCP). World Bank Group, 2015.


125


