Causal Inference with Machine Learning- Examining the Performance of Causal Trees for Analysing Heterogenous Treatment Effects

Research Master Thesis (30 EC)

Moritz Hörl

21/11/2023

Supervisor: Dr. Christopher Clarke

Advisor: Dr. Jack Vromen

Wordcount: 19,874

ERASMUS UNIVERSITY ROTTERDAM

Erasmus Institute for Philosophy and Economics (EIPE)

Erasmus University Rotterdam

Acknowledgements

There would be many people to thank, but I want to keep it rather short. First and foremost, I would like to thank my supervisor Christopher Clarke for his invaluable support and super helpful and insightful feedback during the writing process. I really do not want to imagine how the thesis would have turned out without his guidance. I also want to thank Jack Vromen, who provided me with very helpful comments on an earlier draft of this thesis. In addition, I want to thank Conrad Heilmann for supporting me during my time at EIPE as a tutor with his advice and guidance.

Contents

List of Figures

1. Introduction

The examination and establishment of causal relationships is often at the heart of econometric analysis (Imbens 2022). For example, researchers may be interested how obtaining a microcredit affect the future income of individuals. In such a context, the predominant number of studies and authors is concerned with estimating the average treatment effect (Angrist and Pischke 2009). In essence, this entails to examine if the average income decreases or increases following the receipt of a microcredit.

However, estimating the average treatment effect is often of limited interest for policymakers as they aim to establish targeted policy interventions for individuals with strong positive treatment effects (Kravitz, Duan, and Braslow 2004). This is the case as treatment effects are likely to vary with distinct individual characteristics. In addition, a positive average treatment effect does not imply that the treatment yields a positive treatment effect for every individual, as people with certain characteristics may benefit over proportionally from a particular policy measure. For instance, individuals with higher levels of education may be better in utilizing microcredits they receive, while those with lower education are actually worse off.

The estimation of heterogenous treatment effects does not only hold a prominent place within the discipline of economics but also plays a crucial role in other disciplines. For example, it is of utmost importance for doctors to predict the effect of a drug for a specific patient since new evidence shows that people respond very differently to the same drug (Kent, Steyerberg, and Van Klaveren 2018). Unfortunately, traditional econometric and statistical methods encounter large difficulties when it comes to estimating and discovering heterogenous treatment effects (Benini and Sperlich 2022). This is the case as these methods often rely on a priori knowledge of the researcher to identify variables which may be responsible for differences in treatment effects (Athey 2018). These variables are then included in the form of interaction terms with the treatment variable, thus indicating if the size of the treatment effect depends on the selected variables. However, the a priori knowledge of selecting these variables is often unavailable for researchers. Consequently, algorithmic methods selecting these variables from a large, predefined set of potential variables associated with heterogenous treatment effects would serve as a useful remedy, mitigating the a priori knowledge burden of scientists examining heterogenous treatment effects. This is the case since researchers would only need to prespecify the larger set of variables, without having to precisely select the variables that presumably entail different treatment effect responses.

Due to the upcoming of larger datafiles and increasing computational power in combination with machine learning methods, many promising algorithmically driven methods for detecting treatment effect heterogeneity have emerged in recent years (Gong et al. 2021). One of the most notable methods is the causal tree methodology, developed by the distinguished econometricians Susan Athey and Guido Imbens (Athey and Imbens 2016). Their objective was to combine traditional econometric causal analysis with the widely employed decision tree approach in machine learning. By doing so, they claim that causal trees exhibit superior performance in capturing heterogenous treatment effects compared to traditional econometric methods. While their method and subsequent extensions of it raised strong interest in the academic community (Athey and Imbens 2019), the causal tree methodology has never been comprehensively scrutinized.

This thesis aims to analyse the causal tree method and discusses its potential to provide scientists not only with a reliable method for estimating heterogenous treatment effects, but also for identifying the underlying reasons for these treatment effect differences. It, therefore, seeks to contribute to the literature on applying algorithmic methods for the analysis of treatment effect settings. The intended audience for this thesis includes scientists aiming to employ the causal tree method, but also researchers who are actively engaged in the adaptation of machine learning methods for the application in economics and the social sciences. Given the diverse target groups, this work will provide comprehensive explanations and introductions to technical concepts. However, it is important to note that while all necessary technical prerequisites for following the general argumentation are outlined in the thesis, some parts will require a deeper econometric understanding.

The main contributions are the following:

1) I introduce causal trees with the help of decision trees and causal graph notation instead of falling back on the conventional potential outcome framework. This approach enables a more effective exposition of the advantages of causal trees in comparison to traditional econometric methods.

2) I distinguish between two possible interpretations of causal trees, ultimately leading to the rejection of causal trees as an adequate method for uncovering the underlying mechanisms and revealing the reasons for differences in treatment effects.

3) I show that causal trees fail to provide scientists with reliable individual treatment effect estimation, which constitutes one of the main goals of causal trees. In addition, I argue that

causal trees are unreliable at estimating average treatment effects within subgroups. This also refutes the potential of the causal tree method as an approach for clustering treatment effect heterogeneity. Subsequently, I reject causal trees as a suitable method to examine treatment effect heterogeneity.

This thesis is structured as follows:

In chapter 2, I introduce the causal tree method with the help of the decision tree algorithm and causal graph notation. Furthermore, the important requirement of unconfoundedness will be discussed. I compare standard econometric methods for detecting treatment effect heterogeneity with causal trees, aiming to unveil the advantages inherent in algorithmically searching for differences in treatment effects. Within this context, I will specifically focus on the issue of a priori knowledge in economic research. In particular, I will argue that algorithmically analysing treatment effect settings alleviates the need for a priori knowledge since researchers can not only include more control variables, but also do not have to prespecify heterogenous treatment effect subgroups themselves. In addition, chapter 2 introduces the concept of causal forests, an extension of the causal tree methodology. However, I will show that causal forests face a trade-off between satisfying the unconfoundedness requirement and the establishment of valid confidence intervals. Therefore, I will claim that causal forests are ill-suited for the analysis of treatment effect scenarios.

In chapter 3, I come up with the distinction between a weak and strong interpretation of causal trees. This distinction is necessary due to the lack of one consistent interpretation of causal trees in the existing literature. While interpreting causal trees according to a strong interpretation implies that researchers can gain insights into causal relationships, a weak interpretation primarily treats causal trees as a clustering technique.

In chapter 4, I raise three challenges to the causal tree method, aiming to disclose its limitations. Firstly, I introduce the notion of inconsistent variables in causal trees which lead to high individual variance. Consequently, I claim that causal trees are unreliable at estimating individual treatment effects, which is one of the main goals of the approach. Secondly, I focus on the issue of tree instability and show that there are good reasons to believe that causal trees are a highly unstable grouping mechanism. This also entails high individual variance characteristics in causal trees and can impede the reliable estimation of individual treatment effects. Thirdly, I discuss the concept of M-bias within causal trees and argue that the high number of employed control variables to satisfy the requirement of unconfoundedness makes causal tree results susceptible to M-bias. This is the case as the task of checking for M-bias becomes insurmountable for researchers. In addition, I illustrate that the causal tree method is prone to simultaneously encountering M-bias and confounding bias due to the incorporation of a larger number of control variables. Consequently, I argue that causal trees should be considered as an unreliable method for estimating average subgroup treatment effects. Furthermore, in chapter 4 I provide arguments why a strong interpretation of causal trees cannot be adopted.

In the end, I summarize the main contributions of this thesis to the literature on causal trees and provide some suggestions how machine learning methods may be fruitfully included in economics and the social sciences. While this thesis remains rather sceptical about the applicability of machine learning in the analysis of treatment effect settings and thus, causal analysis, there are continuous technical advancements in this field. Consequently, there may be important methodological developments happening in the next years changing that evaluation.

2. From decision trees to causal forests

The following chapter aims to introduce causal trees, a machine learning method developed for analysing treatment effect settings. In order to allow for a better understanding of the causal tree method, I will begin by discussing decision trees, as the causal tree methodology is built upon them. Alongside describing the technical setup of causal trees, potential advantages over standard econometric methods will be scrutinized and assessed. Furthermore, in chapter 2.3, I will discuss the requirement of unconfoundedness for causal trees with the help of causal graph notation developed by Pearl (2009). In addition, the extension of the causal tree method into a related method called causal forests will be introduced. However, since causal forests can only be employed in settings with a small number of variables, I will demonstrate that causal forests are unreliable for estimating treatment effects.

Given that explaining the method of causal trees involves numerous technical concepts, the coming subchapters will mostly follow a similar structure. First, the main idea of the section will be presented in simple terms, avoiding technical definitions. Subsequently, additional background information will be provided, along with the introduction of more technical notions to fully elucidate the methods and underlying mechanisms. As a result, readers who are not inclined towards technical expressions can abstain from delving into the technical details, as the argumentation can be comprehended based on both the simplified and elaborate explanations.

2.1 Decision trees

Decision trees have a long history in machine learning dating back to the early 1960s, when the first decision tree algorithm[1] was developed (Quinlan 1986). Today, decision trees are still highly used in various fields and have undergone multiple extensions and adaptations since their introduction. Not only have the applied algorithms experienced substantial revisions, but also have decision trees been combined with other machine learning techniques such as deep learning and reinforcement learning (Rokach and Maimon 2014). The primary goals of decision trees are prediction and exploratory data analysis, which have remained unchanged over time (Kotsiantis 2013). While an extensive introduction to decision trees would go beyond the scope

[1] Algorithms can be understood as mathematical procedures which help one to make predictions and thus, learn from the data. These procedures are automatized and therefore, can be applied off the shelf. (Athey and Imbens 2019)

of this chapter, the aim is rather to provide the reader with all the necessary background knowledge to follow the provided argumentation in this thesis.

In essence, a decision tree is a graphical representation that illustrates various averages (Breiman et al., 2017). The construction of a decision tree begins with a random representative sample from the population of interest, comprising multiple data points (Quinlan 1986). These datapoints consist of a dependent variable (e.g. income) and various characteristics, also known as predictor variables, such as age and education. Suppose a researcher is interested in predicting the income of an individual outside the sample. Utilizing the overall average income value of the entire sample may not yield accurate predictions due to the diverse nature of the individuals within it. In order to enhance the predictive accuracy, a decision tree aims to group people based on similar characteristics. For example, individuals may be grouped based on their educational background. By calculating the average income of people with the same educational background, a better predictive performance can be achieved since people with a similar educational background tend to have similar income levels. For example, there is ample evidence that people with a university degree have higher incomes on average than people who quit school relatively early (Oh, Ra, and Jee 2019).

Figure 1 depicts a decision tree that generates subgroups based on age, education and motivation to better predict the income of an individual. It is essential to consider a decision tree from top to bottom. For example, the entire sample in figure 1 is split into two subsamples based on the age of the individuals in the sample. More specifically, people below and above 50 are grouped into two subsamples. Examining the decision tree further, the two subsamples based on age are subsequently split according to educational level and their motivation. However, it is crucial to note that the second splits are contingent upon the first split, signifying that the grouping process continues from the two subsamples of individuals above and below the age of 50. Ultimately, the sample has been grouped into eight different subgroups based on specific characteristics[2]. Finally, the average income of all individuals within each subgroup is calculated. When predicting the income value of a random individual outside the sample, the person is assigned to one of the eight subgroups based on her characteristics, and the average income value of that group is utilized as the prediction for her income. Importantly, the predictive accuracy increases as more characteristics of the person of interest are taken into consideration while progressing down the tree. In addition, the selection of characteristics for

[2] It should be noted that the third splits (high motivation) are the same for each subgroup for illustrative purposes.

grouping the sample into the eight subgroups is not predetermined but carried out by the decision tree algorithm automatically. Consequently, the decision tree algorithm is an algorithm for deciding where the splits should be and thus, for generating a decision tree. Therefore, decision trees enable more accurate predictions of a dependent variable, such as income, by generating numerous subgroups based on various characteristics that play a crucial role in predicting an individual's income level (Breiman et al. 2017).

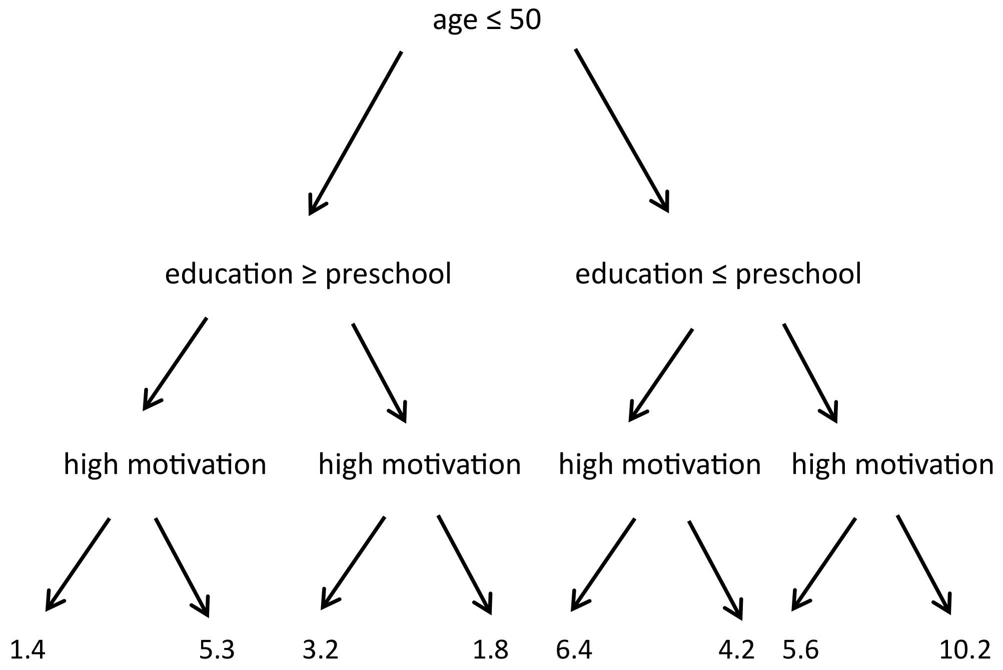

Figure 1. Decision tree predicting the income level of individuals.

While similar predictions could be made with simple linear regression methods[3], decisions trees are more flexible since they do not assume linear relationships between the predictor variables (Kotsiantis 2013). In a linear regression framework, the relationship between the outcome (income) on the one hand and each of the predictor variables (age, education and motivation) is assumed to be linear (James et al. 2013). Even though this method can be effective in many applications, real world systems may not always adhere to linear relationships. While linear regressions cannot account for non-linear relationships between the predictor variables, decision trees can equally well model non-linear relationships between variables (Kotsiantis 2013). In addition, since decision trees can incorporate many variables with different relationships between them, one speaks of high-dimensional decision trees (Athey and Imbens 2019). Despite different possible understandings of the term high dimensionality, in this thesis

[3] Linear regression methods make use of linear functions to depict the relationship between different variables and quantify their relations (James et al. 2013).

it is understood as the possibility to include many different variables into the model with potentially non-linear relationships. While in the standard linear regression case only a few variables can be employed to predict the income of an individual without the method becoming unreliable, researchers can arguably include many more variables applying high dimensional methods like decision trees (Quinlan 1986).

In linear regressions, all variables included in the model are used to predict the outcome (James et al. 2013). In contrast, decisions trees may only use a subset of the included variables to construct a decision tree similar to that in figure 1. The decision which variables are used for predicting the outcome and generating the splits are automatically determined by the algorithm (Quinlan 1986). This is important as it allows for my distinction between *active predictor variables* and *non-active predictor variables* in decision trees, which will be of importance later in this thesis. While *active predictor variables* are actually employed to generate different subgroups in the decision tree estimation procedure, *non-active predictor variables* are included as potential predictor variables by the researcher but not selected by the algorithm to split on. In other words, the decision tree algorithm automatically selects a subset of variables (*active predictor variables*) from the set of all included predictor variables for generating the splits, while the rest of the included variables (*non-active predictor variables)* are not used for building the decision tree. In addition, the researcher can only determine the set of all included predictor variables before employing the decision tree algorithm but does not have any influence on which variables the decision tree algorithm performs the splitting procedure.

As has been mentioned earlier, a decision tree aims to create different subgroups based on some characteristics to predict the value of some outcome variable. In order to do so, decision trees employ splitting criteria. Splitting criteria can be understood as optimization tasks that allow decision tree algorithms[4] to detect the best splits for predictive performance (Athey 2018). One of the most commonly used splitting criterion is the mean squared error (MSE) (Breiman et al. 2017). The objective of this splitting criterion is to minimize the difference between the actual value of the outcome vs the predicted value of the outcome at each splitting point (Quinlan 1986). In the context of the example depicted in Figure 1, this implies that the difference between the actual income value of each individual and the income value predicted by the

[4] The decision tree algorithm determines the process of finding the best splits and building the decision tree. Thus, different decision tree algorithms lead to different decision trees.

decision tree should be minimized as much as possible. This entails the following optimization criterion for splitting decisions:

$$\min_{j,s}\left[\sum_{x_i\in R_1(j,s)}(y_i-\bar{y}_1(j,s))^2+\sum_{x_i\in R_2(j,s)}(y_i-\bar{y}_2(j,s))^2\right]$$

In the case of the MSE criterion, the algorithm searches for the minimal squared difference between the observed and predicted values (Breiman et al. 2017). Again, in terms of example 1, the algorithm tries to minimize the difference between the actual income values and the income values predicted by the decision tree. Once the result is obtained, the algorithm divides the data into two subgroups based on the lowest difference between actual and predicted values and the process repeats (Breiman et al. 2017). Basically, at every point the decision tree splits, new subgroups are created according to some criterion like the MSE. This process continues until a predetermined stopping point. This is necessary as the decision tree would grow infinitely large without any stopping point, deteriorating the interpretability of the resulting tree (Quinlan 1986). While the machine learning literature proposes many possible stopping rules (James et al. 2013), it is mostly set to some predictive performance threshold. Consequently, the decision tree stops generating new subgroups when adding further splits only marginally improves the predictive performance (Breiman et al. 2017). For instance, the decision tree in figure 1 splits only once on age instead of creating more and finer subgroups with respect to age. The decision may be driven by the fact that the predictive performance is already very strong, rendering the creation of additional subgroups unnecessary.

In general, finding the right stopping point is challenging. While generating too many subgroups deteriorates the interpretability of decision tree results, it also may lead to the phenomenon of overfitting (Quinlan 1986). In the case of overfitting, the decision tree contains too many splits and as a result, fits too much noise or random fluctuations to the model (Breiman et al. 2017). Because all observations potentially involve random fluctuations, grouping data into larger groups can be an effective strategy to avoid being led astray by these random fluctuations. Larger subgroups encompass more observations, which facilitates the balancing and reduction of the potential influence of random fluctuations on the predictive value of these subgroups. Therefore, a trade-off exists between achieving excellent predictive performance within the sample and obtaining generalizable results that extend beyond the sample (Athey and Imbens 2019). Although a decision tree might perform exceptionally well for the data at hand, it may exhibit poor predictive performance for out-of-sample units due to overfitting (sensitivity to random fluctuations). Importantly, researchers are in most cases not interested in capturing specific data sample characteristics but aim for generalizable results (James et al. 2013). Therefore, finding the right stopping point is of utmost importance for finding the right balance between including sample related characteristics and aiming for generalizable results.

Due to the elaborated reasons, reducing overfitting and thereby improving the general performance of the model is crucial for decision trees. In order to achieve this goal, multiple different techniques like pruning and cross-validation are applied to decision trees (Breiman et al. 2017). Simply expressed, these methods remove splitting points from the model that do not improve its predictive accuracy (Quinlan 1986). Thus, every splitting point is individually evaluated with additional data to determine if it improves the predictive performance of the decision tree. If not, it is simply removed.

One of the most employed methods in this context is reduced-error pruning, which evaluates the effect of the removal on predictive accuracy using a validation set (James et al. 2013). The validation set comprises data points that were not utilized in the building stage of the decision tree but can then be used to evaluate the predictions made by the decision tree (Breiman et al. 2017). Consequently, the original dataset is divided into a training and validation set to later assess the performance of the decision tree during the pruning stage. While the training set is used by the decision tree algorithm to build the decision tree, the validation set is only used to evaluate the predictive performance of the decision tree. If removing a splitting point improves the predictive accuracy of the regression tree on the validation set, the node is pruned. In general, the pruning stage continues until further pruning does not entail higher predictive performance on the validation set (Breiman et al. 2017). Finally, the resulting decision tree can be assessed based on accuracy and predictive performance using established measures like the R-squared score or additional cross-validation techniques, that partitions the data in several testing subsets (James et al. 2013). However, this step is optional and thus, does not require further elaborations.

In addition to the before explained pruning method, several extensions modifying the underlying decision tree method have been proposed in the literature to resolve the issue of overfitting. The most important one, often referred to as bagging, was introduced by Breiman (1996). Essentially, bagging combines many decision trees and takes the average value of the individual decision tree predictions. Due to technical reasons, the individual decision trees do not need to apply pruning strategies[5]. The most popular bagging strategy is the random forest, which consists of multiple averaged decision trees. In a nutshell, various decision trees are estimated, their predictive values for each individual are aggregated by summing them up and divided by the total number of generated decision trees (Breiman et al. 2017). Given its powerful features, the random forest algorithm has gained widespread popularity in machine learning (Hastie, Tibshirani, and Friedman 2009). One drawback concerning the application of bagging algorithms concerns its difficulty to interpret the results. In contrast to decision trees, which are relatively easy to visualize and understand, bagging methods like random forests cannot be visualized anymore as it is difficult to depict many averaged decision trees (Strobl et al. 2007).

While developers in the field of machine learning primarily focus their attention on addressing the issue of overfitting, econometricians are more interested in the variance and bias properties of these algorithms (Breiman 2001). Since econometricians try to avoid both high variance and biased results, a trade-off arises between these two properties.[6] In general, when econometricians discuss the variance of an estimator, they are concerned with what will be referred to as the *sample variance* in this thesis. To better clarify what is meant by the term *sample variance*, it may be helpful to think about the following experimental setting:

Consider a scenario where researchers are interested in the average height of children at the age of 8 in the Netherlands. Ideally, the researchers would be able to measure every single 8-year-old child in the Netherlands, thereby obtaining the true value with absolute certainty. However, this appears impracticable due to the associated costs and logistical difficulties. Instead, they will take a representative sample of children at the age of 8 in the Netherlands and measure their average height. Consequently, the measured average height is then taken as the best approximation of the true average height of children at the age of 8 in the Netherlands. Nevertheless, there will persist some measurement uncertainty since the researchers have not measured every 8-year-old child in the Netherlands but have taken a sample. This source of uncertainty can be understood as *sample variance*. Since researchers are interested in the precise population value, they try to avoid high *sample variance* results. One way to decrease the

[5] For interested readers: Pruning is not necessary since errors cancel out when the different decision tree predictors are combined to reduce overfitting (Breiman 1996).

[6] The bias-variance trade-off states that the more variables are included into a model, the lower the bias as the fit of the model increases. However, adding more variables increases the variance of the model. (Angrist and Pischke 2009)

uncertainty and hence the *sample variance*, involves increasing the sample size. This appears quite intuitive as measuring the height of more 8-year-old children would provide the researchers with more information about the true average height of children at the age of 8 in the Netherlands and thus, decreases the *sample variance* (uncertainty).

Similar to high *sample variance*[7] outcomes, researchers seek to avoid biased results (Angrist and Pischke 2009). Bias occurs when some factors distort the outcome of a model, leading to either systematic over- or underpredictions of the value of interest[8] (Athey and Imbens 2017). Referring back to the previous example, biased results would render height level predictions unreliable. A potential reason for the emergence of such bias could be the use of skewed height measurement scales for measuring the height, leading to measurement errors.

In contrast to the *sample variance* of a result, researchers can also be interested in the *individual variance* of a result. *Individual variance* in this thesis is understood as a measure of reliability of the estimated result when applied as a prediction for an individual not included in the original sample. Referring back to the example from above, the *individual variance* indicates how well the measured average height from the selected representative sample can be taken as an approximation for the height of an eight-year-old child in the Netherlands outside the sample. For example, if the children´s heights in the sample significantly differ, the average height may serve as a poor predictor for the height of a randomly selected eight-year-old child in the Netherlands. However, in cases where the variability in children´s height within the population is minimal, taking the average height as a prediction for a child outside the sample may be appropriate. Consequently, *sample variance* and *individual varianc*e are not necessarily related. While the *sample variance* may be very low in cases involving a large sample, the *individual variance* may be high due to a significant height variability within the examined population. This distinction is important, for example in contexts such as assessing the impact of a drug on the survival rates of cancer patients. In such scenarios, it is not only crucial to have a high reliability of the true value of the drug´s average effect in the population (low *sample variance*), but also to have knowledge of the *individual variance* of the drug´s effect. This knowledge allows doctors to better predict the potential effect of the drug on an arbitrary individual from

[7] High *sample variance* results are often a good indicator for overfitting tendencies. This is the case since models with *high sample variance* tend to include many variables and hence, capture many specific data peculiarities. Consequently, like in the case of overfitting, the model entails a bad out of sample performance.

[8] One reason for biased results will be extensively discussed in chapter 2.3.

the population and could, in case of high *individual variance,* prompt additional considerations and adjustments.

The key takeaway from this chapter is that decision trees are a powerful method to predict outcomes in settings with many variables (characteristics) and potentially non-linear relationships between them. To achieve this, decision trees apply an optimization criterion and generate different subgroups as was presented with the help of figure 1. Nevertheless, it should be noted that decision trees entail the issue of overfitting that makes the application of pruning or extensions like random forests necessary. In addition, the distinction between *sample variance* and *individual variance* of an estimation result was drawn and discussed as an understanding of these concepts is important for the further argumentation in this thesis.

2.2 Causal trees

In comparison to decision trees, causal trees represent a recent development and are primarily employed in treatment effect settings (Imbens 2022). For instance, researchers may use them to better analyse the effect of a drug on the recovery time of people having the flu. In essence, causal trees exhibit many similarities with decision trees. Similar to decision trees, causal trees aim to establish heterogeneous subgroups based on characteristics (splitting variables) (Athey and Imbens 2016). Although the two algorithms slightly differ concerning their technical aspects[9], the underlying mechanism is similar. In general, the causal tree algorithm differs from decision trees in only two ways:

First, causal trees estimate treatment effects rather than focusing solely on predictive outcomes (Athey and Imbens 2019). In other words, causal trees are used to analyse treatment effect settings, which aim to investigate the effect of a treatment on a particular outcome. For instance, a typical application of causal trees is to examine the effect of receiving a microcredit on the future income of an individual. Therefore, the average treatment effect for all established subgroups is estimated and then employed as an approximation of the individual treatment effect for individuals outside the sample. In other words, the average treatment value of the subgroup to which an individual would belong is taken as the estimate of the individual treatment effect for that individual. Given that the causal tree algorithm does not provide researchers with exact individual treatment effect estimations, they must rely on average subgroup treatment estimates. Establishing subgroups with different treatment effects allows,

[9] Since the technical details are not important for the further argumentation provided in this thesis, I will not present them here. Instead, I refer interested readers to Athey and Imbens (2016).

for example, doctors to provide patients with tailored drug recommendations instead of following generalized guidelines designed for the average patient across all subgroups. Consequently, causal trees appear to be a powerful and valuable method for exploring heterogeneity in treatment effects as it establishes several subgroups with different treatment effects.

Second, the causal tree algorithm employs different segments of the sample to select the splitting points and estimate the treatment effects for the specific subgroups (Athey and Imbens 2016). Initially, the training set is divided into two parts. While half of the training dataset is used in the estimation process to determine the variables on which the causal tree is splitting, the other half is employed to estimate the treatment effect for each established subgroup. Hence, the causal tree algorithm can be imagined as a three-step process: First, subgroups are formed based on one part of the training sample data. Second, another part of the training sample data is used to estimate the average treatment effect for each of the established subgroups. Third, some subgroups are removed through cross-validation and pruning techniques to prevent overfitting, employing the cross validation set. This process is in the literature referred to as *honest estimation* (Athey and Imbens 2016). In contrast, the decision tree algorithm is simultaneously establishing different subgroups and making predictions for each subgroup, and can therefore, only be described as a two-step process (Breiman et al. 2017). According to causal tree proponents, the additional step in their methodology is necessary for obtaining confidence intervals for the treatment effect results (Athey 2018). Confidence intervals serve as a metric for assessing the reliability of the causal tree results, providing a range within which reliable outcomes lie (James et al. 2013). For instance, if a 90% confidence interval is constructed for the estimated income value of a subgroup, it implies that there is only a 10% chance that the true income value[10] lies outside the confidence interval.

Moreover, results with smaller confidence intervals are more reliable than results with larger confidence intervals. This is the case as smaller confidence intervals imply less variation and consequently, reduced uncertainty in the predicted income value (Hastie, Tibshirani, and Friedman 2009). Therefore, the reliability of a result increases as the confidence interval becomes narrower. As confidence intervals serve as a measure of reliability, they are especially important when translating estimation outcomes into policy recommendations. Therefore, the

[10] The true income value describes the unknown true value one tries to estimate with the help of statistical methods.

adapted causal tree methodology with *honest estimation* represents a crucial step in enhancing the applicability of tree algorithms, as it allows for the analysis of result reliability.

2.3 Unconfoundedness in causal trees

One important assumption for the application and a proper understanding of causal trees is the requirement of unconfoundedness, which will be introduced and discussed in the section to come. While the requirement of unconfoundedness can be presented using various frameworks, I will focus in this section on Pearl´s causal graph notation and argumentation as presented in Pearl (2009) due to its intuitive nature and ease of visualization.

As has been elaborated in the previous section, causal trees are primarily employed in treatment effect settings. Therefore, researchers are interested in examining the effect of a treatment on some outcome variable. The unconfoundedness requirement now posits that the outcome is only influenced by the treatment and other variables that are not correlated with the treatment (Angrist and Pischke 2009). This leads to the result that the differences in outcomes between units receiving the treatment and units not receiving the treatment can be exclusively attributed to the treatment itself.

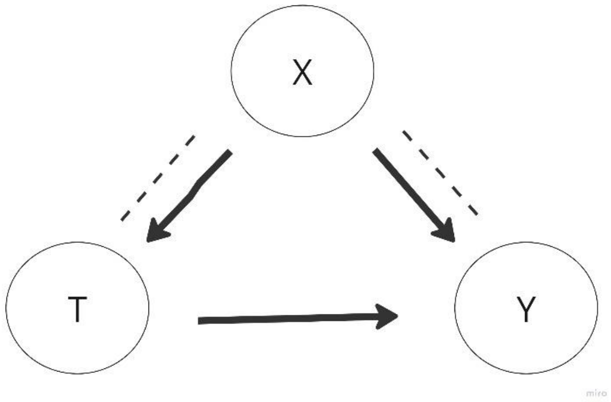

Figure 2. Unconfoundedness assumption in causal graph notation.

In terms of the causal graph notation, the requirement of unconfoundedness can be expressed as the requirement that all causal backdoor paths are closed (Pearl 2009). Let me now explain what this means in more detail. In general, there are three ways variables can be associated[11] with each other. In other words, there are three possibilities for an open causal path between two variables. First, there can be a direct causal relation between two variables. As depicted in figure 2, T (treatment variable) has a direct causal impact on Y (outcome variable), indicated

[11] It is important to note that in this thesis association and correlation are used interchangeable.

by the arrow from T to Y. In addition, there is a second option how two variables can be associated. Two variables can share a common cause, which opens a non-causal association between the two. As depicted in figure 2, X is a common cause of both variables, T and Y. This is shown by the two arrows, directed from X to T and Y. Consequently, figure 2 depicts the situation that X has a causal effect on T and Y. However, such a situation is problematic for analysing the effect of T on Y since the two variables are not only associated through the direct causal path from T to Y, but also through a non-causal association via the common cause X. Consequently, using the correlation between T and Y to estimate the effect of T on Y would yield biased results, as both the causal and non-causal paths would contribute to the correlation between X and Y and thereby to the estimate of the causal contribution X makes to Y. Thus, the common cause X is acting as a confounding variable, affecting the estimated effect of T on Y, and leading to biased results due to the influence of the non-causal link between T and Y via X. Furthermore, there exists a third option how two variables can be associated with each other. More specifically, two variables can share a common effect, resulting in the emergence of a collider. In contrast to the scenario illustrated in figure 2, a collider situation would imply that the causal arrows are leading from T and Y to X. Consequently, both variables T and Y would cause the variable X. This situation can pose problems for the estimation of causal effects as it again opens an additional non-causal association between T and Y, similar to the second situation in which X functioned as a confounding variable.

When analysing treatment effect settings, researchers are primarily interested in the causal effect of T on Y. Consequently, researchers need to be cautious in opening and blocking causal paths. In general, a causal path can be understood as an association between two variables as depicted in the causal graph notation by an arrow leading from one variable to the other. Given that the effect of Y on T is the subject of interest, researchers need to make sure to block all non-causal associations possibly emerging as the result of a common cause or common effect. However, there are important differences between blocking the causal path of a common effect and a common cause:

To block the non-causal path between T and Y via X in the case of a common cause (potential confounding variable), the common cause X can be included as a "control variable" in the estimation process (Cinelli, Forney, and Pearl 2020). Thus, the non-causal association between T and Y can be blocked through employing the potential confounding factor as a variable in the estimation process. Control variables can be analogously understood as predictor variables in the case of a decision tree, only that one speaks of control variables in a treatment estimation framework. In contrast, a non-causal association in the case of a common effect emerges only when researchers include a control variable, which happens to be a common effect of T and Y, in the estimation process. Importantly, the association between T and Y is dependent on the stratification of the common effect control variable. In other words, the non-causal association between Y and T only arises when researchers control for the specific variable in the estimation process (employ it as a control variable in the estimation process). If one does not control for the common effect variable, T and Y are independent of each other[12]. Researchers often speak of stratified correlation in this context (Pearl 2009). Consequently, while the issue of common cause variables (confounding variables) can be addressed by employing the specific factor in the estimation process, it is the other way around in the case of a common effect. By default, T and Y are independent of each other. An association only arises if the common effect factor is introduced as a control variable, connecting T and Y in further consequence. In the context of causal trees, controlling for variable X simply means that one allows the causal tree to split on X if it improves the estimation of treatment effects. Consequently, the variable X is included as a control variable in the causal tree estimation process in case the correlations would give rise to a confounding situation[13]. As open non-causal associations lead to biased and unreliable results, it is crucial for the causal tree method to identify potential confounding variables and incorporate them as control variables into the causal tree estimation process, despite the risk to control for a common effect variable. An example of a potential confounding variable can be illustrated with the following situation:

Suppose a researcher aims to examine the effect of microcredits on future income. In this example, the treatment variable (T) indicates whether an individual receives the microcredit or not, and the outcome variable (Y) is the future income. In addition, the researcher introduces age as a control variable (X) as she believes that age may potentially be a confounding variable. On the one hand, the researchers knows that certain age groups are more likely to receive the treatment. On the other hand, she is convinced that age affects future income since younger individuals tend to be more diligent than older ones. Thus, including age as a control variable becomes necessary to block the non-causal association between the treatment and outcome

[12] This of course only holds true when the two variables are not associated with each other through a confounding variable or direct causal path.

[13] Although the causal tree has the option to split on X, it does not have to be the case as it may not necessarily improve the estimation of the average treatment effects for the respective subgroups. In addition, this cannot be decided by the researcher but is automatically implemented by the tree algorithm as described in chapter 2.1.

variable. Without the inclusion of age as a control variable, the resulting estimates would be significantly biased.

However, identifying potential confounding variables is not a straightforward task and often involves contentious debates. This is the case as confounding variables cannot be detected through statistical testing but only through theoretical reasoning. In other words, the assertion that one variable caused another one cannot be tested empirically, without relying on prior knowledge, theory, or intuition. While it is possible to test for correlative associations between variables (Hastie, Tibshirani, and Friedman 2009), the same does not hold true for causation. Going back to the example mentioned earlier, it may be equally plausible that the researcher is mistaken about the causal link between age and future income. Nevertheless, such a proposition cannot be empirically tested but only argued for on a theoretical level. In addition, controlling for a potential confounding variable may yield unintended consequences as controlling for a variable may result in a collider association in case the control variable is the common effect of T and Y. While confounding bias occurs when researchers fail to control for a specific variable, bias stemming from a collider variable only arises if one controls for a common effect variable. Consequently, it is of utmost importance to have strong theoretical reasons to believe that a variable is a common cause (confounding variable) between the treatment and outcome variable when employing it as a control variable.

Consequently, in standard econometric methods the researcher's beliefs play a crucial role in deciding which variables to include as control variables. As a result, two researchers analysing the same treatment effect setting may yield different results since they might include different control variables, leading to different estimation results. The causal graph notation introduced in this chapter provides researchers with the possibility to visually examine the unconfoundedness requirement, which states that researchers need to control for every potential confounding variable to block the non-causal association between the treatment (T) and outcome variable (Y).

2.4 Advantages of causal trees over standard econometric methods

After pointing out the differences between decision trees and causal trees and introducing the important requirement of unconfoundedness, this section aims to address some prima facie reasons why economists and social scientists should adopt the causal tree method. Specifically, I want to emphasize two prima facie advantages of causal trees over standard econometric methods in their analyses of treatment effect settings. While causal tree proponents often focus on the technical details of the causal tree algorithm, less attention is devoted to discussing issues related to the practical implementation of causal trees. Consequently, I try to fill this gap in the literature by systematically examining the advantages of the causal tree method over standard econometric techniques. First, I will claim that prima facie causal trees yield more reliable results since they better satisfy the aforementioned requirement of unconfoundedness. This is due to the fact that causal trees can incorporate a greater number of control variables compared to standard econometric methods. Second, I will argue that causal trees do not rely on theoretical knowledge for forming specific heterogenous treatment subgroups and are hence superior from an epistemic standpoint.

Causal trees appear to be advantageous over standard econometric methods as they allow to include more control variables in the estimation process. In social science settings, there are numerous potential confounding variables, which often cannot be known a priori by the researcher (Keane 2010). Due to their large number, one cannot incorporate all potential confounders as control variables in a regression equation in the standard econometric framework as this would inevitably lead to overfitting (Pacifico 2021)[14]. The underlying reason is similar to the cause for overfitting in decision trees as has been discussed in chapter 2.1. In case of incorporating too many variables, the resulting model tends to entail bad generalizing properties since it is well fitted to the noise in the sample data but performs poorly on out of sample data.

In contrast, machine learning methods like causal trees can handle large sets of control variables as they are designed to operate in settings with a high number of variables (Breiman et al. 2017). The difference to standard econometric methods lies in the fact that the tree algorithm does not have to employ them as splitting variables (and hence incorporate them in the estimation process). Instead, they can be incorporated in the control variable pool as potential variables to

[14] For a more detailed explanation of overfitting and the variance-bias trade-off see chapter 2.1.

split on.[15] This resembles the distinction drawn in chapter 2.1 between *active predictor* and *non-active predictor* splitting variables. On the other hand, standard econometric methods require all employed variables to be included in the estimation process. Hence, standard econometric techniques solely incorporate *active predictor* variables. Consequently, situations like the one depicted in figure 3 can only be effectively analysed using machine learning methods. A potential real-world example illustrating this issue can be taken from the microcredit literature:

Measuring the effect of microcredits (the treatment) on income is a situation with many potential confounding variables. Not only can individual characteristics such as age, education, risk attitude, motivation, and more, be associated with selection into treatment, but also influence the outcome. In addition, there may be various additional potential confounders related to the environmental context, such as geographical location, economic conditions, specific laws/regulations and so forth. As previously mentioned, it is inadvisable to include all potential confounders as control variables in a standard econometrics setting as the risk of overfitting increases with every additionally included variable. Therefore, treatment effect settings which require the inclusion of many control variables can only be analysed using causal trees. Standard econometric methods can only employ few control variables and hence, are susceptible to confounding bias in situations which would require the incorporation of many control variables. Consequently, causal trees are prima facie more reliable in satisfying the unconfoundedness requirement given that numerous potential confounding variables can be included in the estimation process and thus, more non-causal paths be blocked.

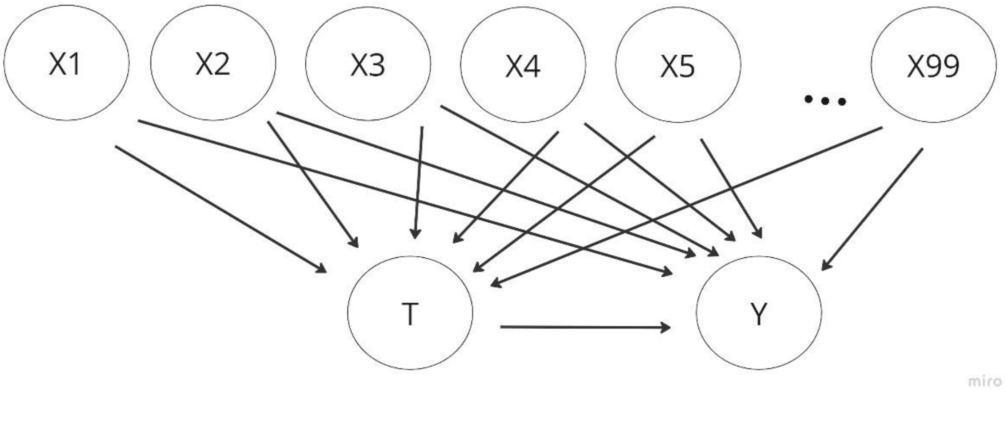

Figure 3. Unconfoundedness requirement only fulfilled through conditioning on many covariates.

[15] Consequently, the causal tree algorithm automatically selects the control variables entailing the highest treatment effect differences.

In addition, causal trees also appear advantageous in relation to forming heterogenous treatment effect subgroups. As discussed in the previous chapter, causal trees classify people into different subgroups and estimate the average treatment effects of the established subgroups (Athey and Imbens 2019). Since the established subgroups differ in their characteristics, the average treatment effect will differ among these subgroups (similar to how predictions vary between different subgroups for decision trees). Therefore, one speaks of heterogenous treatment effects.

In general, the approach to detect heterogenous treatment effects of standard econometric methods and causal trees differs in significant aspects. While heterogenous subgroups must be specified a priori by economic theory or intuition in a standard econometric setting (Fink, McConnell, and Vollmer 2014), causal trees generate them in a data-driven way without prior specification (Athey 2018). In other words, standard econometric methods assume that researchers know how to form the relevant subgroups before even estimating the econometric model. In a standard econometric framework, the subgroups must be defined based on economic theory or established through intuition. In contrast, applying the causal tree method solely requires the specification of all control variables that might serve as potential confounding variables. Consequently, the causal tree algorithm automatically selects from all these many potential confounding variables a smaller set from which to generate the specific subgroups. Since this may appear somewhat abstract to some readers, the following example illustrates this point:

Suppose researchers want to analyse the effect of a microcredit on future income. Furthermore, assume that researchers are interested in the different treatment responses between various subgroups. Analysing it with standard econometric methods requires the researcher to predefine the subgroups before estimating the model. This necessitates careful consideration of certain characteristics that might influence the treatment effect response. For example, the age and education of a person could be factors leading to differences in treatment effects. Younger and well-educated individuals may be better in utilizing microcredits to increase their income. Therefore, in a standard econometric framework, young and well-educated individuals would form a subgroup as researchers expect to obtain different treatment effect results subgroups defined by age and education. This process would be repeated until the researcher is satisfied with how the sample is divided into distinct treatment effect subgroups, and then the average treatment effect is estimated for each subgroup. In contrast, when using causal trees, researchers only need to specify the control variables they want to include, and the algorithm itself determines the subgroups that entail the highest treatment effect differences. This is the case since causal trees only split on a small number of the control variables.[16] Therefore, no economic theory or intuition is required when specifying the subgroups for employing the causal tree method. The subgroups are automatically identified by the algorithm, making it a more data-driven and less theory dependent approach.

As the discussion and example revealed, applying causal trees involves less reliance on economic theory and intuition. While standard econometric methods heavily depend on economic theory to predefine potential heterogenous treatment effect subgroups, algorithmic methods like causal trees employ the available information in the data to generate these subgroups. This aspect can be viewed as a big advantage of causal trees over standard econometric methods since the prior beliefs[17] of economists and social scientists have less influence on the results, given that identifying the subgroups is carried out algorithmically. Consequently, this methodology can be evaluated as more explorative, since the algorithmically established subgroups may serve as a starting point for further research to understand why the causal tree split on specific characteristics to analyse treatment effect heterogeneity. In contrast, the subgroups in a standard econometric framework are predefined by the respective researcher, limiting the potential for further exploratory research.

To express this point in more precise technical language[18], researchers can only identify heterogenous treatment effects in standard econometric settings through the introduction of interaction terms between the treatment and the respective variables which they believe are the drivers of differences in treatment effects (Fink, McConnell, and Vollmer 2014). Therefore, only a few interaction terms can be tested for statistical significance at once since testing multiple hypotheses makes the model susceptible to wrongly ascribing statistical significance to normally non-significant interaction terms[19] (Hoover 2013). Thus, researchers need to be cautious in prespecifying the potential subgroups they are interested in. As a result, the approach is more confirmatory than explorative since not all possible interactions between variables can be included and tested. This is again the case since the a priori beliefs of the involved researchers are important for specifying the subgroups. Consequently, subgroups are only included in the

[16] As previously defined, these variables are called *active predictor* variables in this thesis.

[17] The process of establishing subgroups based on economic theory and intuition is a subjective endeavour with a large leeway for deviations in the grouping process. For example, some researchers may be convinced that age does not have an effect on treatment effect responses and thus, solely define subgroups based on education.

[18] As the main idea has been thoroughly discussed in the paragraph above, readers without appetite for technical language can simply skip this section.

[19] This is the case as each interaction term has the probability of 0.05 (p-value) to be wrongly included in the model.

analysis if researchers already assume them to reveal different treatment effects. This naturally restricts the leeway for scientific discoveries in relation to heterogenous treatment effects. In contrast, as I have shown in this chapter, causal trees establish subgroups in a data-driven way without strong reliance on economic theory.

2.5 Causal forests

Since causal trees are equally prone to overfitting as decision trees (see chapter 2.1), econometricians have employed bagging strategies to construct causal forests. Based on the causal tree methodology introduced earlier, Athey and Wager (2018) proposed the averaging of causal trees to create causal forests similar to the process of generating random forests from decision trees. Even though the algorithm exhibits slight differences, the process of generating causal forests is identical to generating random forests.

Essentially, multiple causal trees are estimated from different subsets of the sample and the average treatment value for each subgroup is computed (Wager and Athey 2018). Due to the same technical reasons as in the case of decision trees, no pruning strategies need to be applied for estimating the causal trees before averaging them[20]. Consequently, causal forests employ large causal trees as additional splits do not increase the risk for overfitting, which would be the case for individual causal trees.

While causal trees estimate individual treatment effects through taking the average treatment effect of established subgroups, causal forests estimate treatment effects for each individual (Wager and Athey 2018). This means that individuals are not grouped into different subgroups but that the causal forest produces individual estimation results for every individual in the sample. This is possible as causal forests can incorporate infinite splits due to the fact that bagging rather than pruning techniques are employed to avoid overfitting. Consequently, causal forests can group people into one-unit subgroups, making the estimation of individual treatment effects feasible. In other words, due to the utilization of large causal trees in the causal forest approach, researchers can obtain individual treatment effect results for every individual.

Nevertheless, interpreting causal forests is difficult as they cannot be easily visualized[21]. Therefore, Athey and Imbens (2019) argue that causal forests should not be seen as a replacement for causal trees. Instead, researchers should select the appropriate method based on the specific estimation problem they encounter. For instance, if doctors aim to determine the

[20] Interested reader can go back to chapter 2.1 for more detailed explanation.

[21] This is again very similar to random forests.

potential treatment effect for a specific patient, causal forests may be a more suitable choice, while causal trees might prevail in situations where a straightforward visualization of treatment outcomes is necessary.

However, causal forests only possess accurate confidence intervals in settings with few included control variables as has been shown by Chernozhukov et al. (2018). This implies that confidence intervals can only be reliably estimated when a small number of control variables are included in the causal forest estimation process. The technical reason for that limitation is that Wager and Athey (2018) hold the dimension in their analysis fixed when proofing the existence of valid confidence intervals for different dimensions[22]. Consequently, the causal forest method does not entail accurate confidence intervals in typical causal tree settings with many included control variables. Hence, the causal forest will have to have only a few included control variables. This characteristic undermines the advantage of the causal tree method presented in chapter 2.4, which stated that causal trees better satisfy the important requirement of unconfoundedness compared to standard econometric methods. The presented argument was that causal trees can incorporate a larger number of control variables and thus, block more potentially non-causal associations arising from confounding. However, as I was arguing in this chapter, causal forests exhibit a trade-off as they either do not allow for the estimation of valid confidence intervals or may only be applicable in settings with a limited number of variables. Both properties are pivotal for obtaining reliable results. First, confidence intervals are needed as a measure of uncertainty of the estimation results. Second, including more control variables is often necessary to satisfy the requirement of unconfoundedness in causal inference settings. Consequently, due to this trade-off, causal forests do not seem to be a suitable estimation method for individual treatment effects.

Interestingly, the issue discussed above has been overlooked in the causal inference literature so far despite major implications for the application of causal forests. Consequently, this thesis is the first work pointing out the dilemma causal forest face: while too many variables make the method unreliable from an econometric perspective, including too few variables calls into doubt the unconfoundedness requirement. Neither is this dilemma specifically brought up in the applied literature on causal forests, nor is it discussed in the theoretical literature on causal

[22] No consistent estimator can be estimated once $d \geqslant \log n$. For a more detailed explanation see Stone (1982).

inference. This is especially surprising given the fact that causal forests have been increasingly applied in the literature over the last years[23].

[23] See for example Davis and Heller (2017), Miller (2020) or Gulen, Jens, and Page (2021).

3. Interpreting causal trees

After introducing the most important technical details and assumptions of the causal tree method, this chapter seeks to distinguish between two approaches to interpreting causal trees, namely a strong and weak interpretation.

3.1 Interpretations of causal trees

In general, it appears that causal trees can be meaningfully interpreted in two different ways. First, causal trees can be seen as a method for understanding the underlying mechanism leading to treatment effect heterogeneity. Since causal trees split on the variables which are best to group the population into different treatment effect subgroups (Athey and Imbens 2016), a strong interpretation of causal trees would assign the splitting variables an explanatory interpretation. Consequently, the variables used for splitting can be regarded as the causal drivers for treatment effect heterogeneity. Taking the formerly used example analysing the effect of microcredits on future income, if a causal tree splits on the control variable age, it implies that age influences the treatment effect of microcredits on future income, attributing a causal interpretation to age. Employing a strong causal tree interpretation, age is considered as a causally contributing factor for the variations in future income.

Second, causal trees can be given a weak interpretation, where the aim is rather to group individuals into different subgroups with varying treatment effects instead of explaining the underlying mechanism responsible for these differences. Therefore, the actual variables used for splitting within the causal tree can be neglected since only the grouping result is of interest. In other words, according to the weak interpretation, causal trees serve as a robust grouping method that categorizes individuals into subgroups with varying treatment effects. The results can then be used to allocate treatments or assign policies. However, evaluating causal trees from this perspective does not allow researchers to draw any conclusions about the reasons for differences in treatment effects. Consequently, the actual grouping process remains a black box since the splits within causal trees cannot be meaningfully interpreted. Thus, no conclusions can be drawn which variables causally affect treatment responses. Applying this definition to the microcredit example, interpreting the causal tree results according to the weak interpretation means that the causal tree method is of help for grouping people into different future income groups but fails in elucidating the underlying reasons for the formation of these divergent income groups. Even though the causal tree may split on age, that does not mean that age is a causal driver for treatment effect heterogeneity.

Based on this distinction, the interpretation of causal trees depends on whether a strong or weak interpretation is adopted. While the former aims to explain the causal tree grouping process, the latter solely focuses on the grouping outcomes without providing a deeper understanding of the differences in treatment effects.

3.2 Strong interpretation of causal trees

Even though causal tree proponents are generally cautious about a strong interpretation of causal trees, some quotes indicate a desire to ascribe causal trees the potential to explain the grouping process. For example, Athey (2018) states: "Treatment effect heterogeneity can be of interest either for basic scientific understanding (that can be used to design new policies or understand mechanisms), or as a means to the end of estimating treatment assignment policies that map from a user's characteristics to a treatment." (524)

While it is commonly stated that causal trees are primarily constructed to detect treatment heterogeneity (Wager and Athey 2018), the aforementioned quote by Athey (2018) implies that causal trees should also contribute to understanding mechanisms. This indicates a strong and causal interpretation of causal trees since understanding a mechanism is only possible if the main driving forces of that mechanism are known. The main driving forces on the other hand can only be inferred from the respective method employed, which, in this case, is the causal tree method. Hence, the requirements of causal trees extend beyond the mapping of "user´s characteristics to a treatment" (Athey 2018, 524) when a strong interpretation is applied. Despite some quotes pointing towards a strong interpretation of causal trees, most causal tree developers are cautious about a causal interpretation of the splitting variables[24]. Nevertheless, social scientists employing causal trees for their research often assume a strong interpretation[25]. This inclination does not come as a surprise given that causal tree developers only provide vague guidelines for interpreting causal trees. As demonstrated in this chapter, causal tree proponents do not consistently apply one interpretation. Consequently, social scientists often adopt a strong interpretation of causal trees since establishing data-driven causal relationships appears more powerful than simply grouping people into different treatment subgroups. Thus, causal tree developers need to engage in further discussions on what the causal tree method can truly accomplish to avoid misunderstandings in its application by other researchers. While the technical aspects of causal trees are extensively discussed in the literature, it is equally

[24] See for example Chernozhukov et al. (2017) and Athey and Imbens (2017).

[25] See for example Bargagli, Stoffi and Gnecco (2020).

important to analyse the two distinct interpretations of causal trees and uncover their respective assumptions.

3.3 Weak interpretation of causal trees

From the perspective of a weak interpretation, the evaluation of causal trees focuses solely on the grouping results, disregarding the splitting variables or the understanding of the underlying mechanism. Even though it is not specifically addressed in the literature, a weak interpretation of causal trees is often assumed to be the underlying objective when estimating causal trees. For instance, Athey and Imbens (2017) state that "One example is to examine within subgroups in cases where eligibility for a government program is determined according to criteria that can be represented in a decision tree,… ." (25). Similarly, the same authors write: "Examples include treatment guidelines to be used by physicians … ." (Athey and Imbens 2016, 7354). These applications do not require a strong interpretation of causal trees but mainly aim to generate subgroups for allocating treatments. Therefore, there is no need for a causal interpretation of the splitting variables. Despite occasional references by causal tree proponents to a strong interpretation, as demonstrated in the previous section, it appears fair to say that the detection of heterogeneous subgroups is one of the primary objectives of causal trees. Consequently, causal trees employed in the academic literature should be primarily interpreted according to a weak interpretation. Furthermore, in section 4.5, I will present additional arguments revealing that upholding a strong interpretation of causal trees is unfeasible.

4 Causal trees unmasked: revealing its limitations

In the following chapter, I will raise three challenges aiming to demonstrate potential pitfalls for causal tree methods on both possible interpretations. The first two challenges are concerned with the problem of high *individual variance* for causal tree estimates, which entail bad individual treatment effect estimations as I will argue in this chapter. In addition, the third challenge is related to the estimation of average treatment effects within established subgroups. More precisely, I will claim that causal trees are susceptible to bias, leading to unreliable results. In this context, unreliability means that the results of the causal tree method cannot be trusted.

Firstly, I will elaborate on the issue of *inconsistent variables* in causal trees and the entailing high *individual variance* properties. Secondly, I discuss causal tree instability and its implications, which similarly contribute to high *individual variance* observed in the estimated results. Thirdly, the notion of M-bias will be introduced and analysed within the Pearl framework.

4.1 Inconsistent variables in causal trees

Statistical methods yielding lower *sample variance* in their results are preferred to those with higher *sample variance*. Since the causal tree method potentially suffers from high *sample variance*, pruning and cross-validation techniques are necessary to decrease the *sample variance* as has been demonstrated in chapter 2.1. While necessary to decrease the *sample variance* of the causal tree method, I want to show in this chapter that pruning may remove important information from the model, deteriorating the estimation of individual treatment effects. While this issue is also important for prediction tasks and hence decision trees, it will be emphasized that causal inference settings are more vulnerable to the adverse effects stemming from aggressive pruning and cross-validation methods due to the application of *honest estimation*. Since this thesis should be accessible to practitioners with non-technical backgrounds, I will discuss the problem of aggressive pruning and cross-validation techniques through the concept of *inconsistent variables*, which will be introduced and analysed in the remainder of the chapter. An *inconsistent variable* can be described as following:

Consider a set of three variables: one serving as the treatment variable, another as the outcome variable, and an additional mediator variable that independently affects the outcome variable. The influence of the mediator variable on the treatment effect can be either monotonic or non-monotonic. In instances where the influence is non-monotonic, the mediator variable is classified an *inconsistent variable*.

To make better sense for the reader what is exactly meant by that understanding, I will provide an example with the help of figure 4, which is the same as was used to introduce decision trees and should now be treated as a causal tree. As illustrated in figure 4, the data is split on age at the top of the tree creating subgroups on the basis of age. However, it is possible that the treatment effect differs not only between people above and below the age of 50 but also significantly between those under the age of 20. Consequently, age does not only act as a mediator variable, but also affects the outcome in the following non-monotonic way: individuals between the ages of 0-20 show a low treatment effect, individuals between the ages of 20-50 on the other hand show a high treatment effect, while the treatment effect is low again for those between the ages of 50-100. Nevertheless, let´s imagine that the causal tree splits once on age as the treatment effect heterogeneity is highest for people below and above the age of 50. Therefore, researchers may draw the deceiving conclusion that the older the person is, the higher or lower the treatment effect. Since the causal tree is not splitting twice on age, it is impossible for researchers to infer this information from the data. In this example, age can be seen as an *inconsistent variable* as the effect size does not monotonically increase or decrease with age but obtains varying values for different age groups. This can result in highly problematic policy implications as people below the age of 20 may not be eligible to apply for microcredits due to their seemingly low treatment effects (as depicted in figure 4).

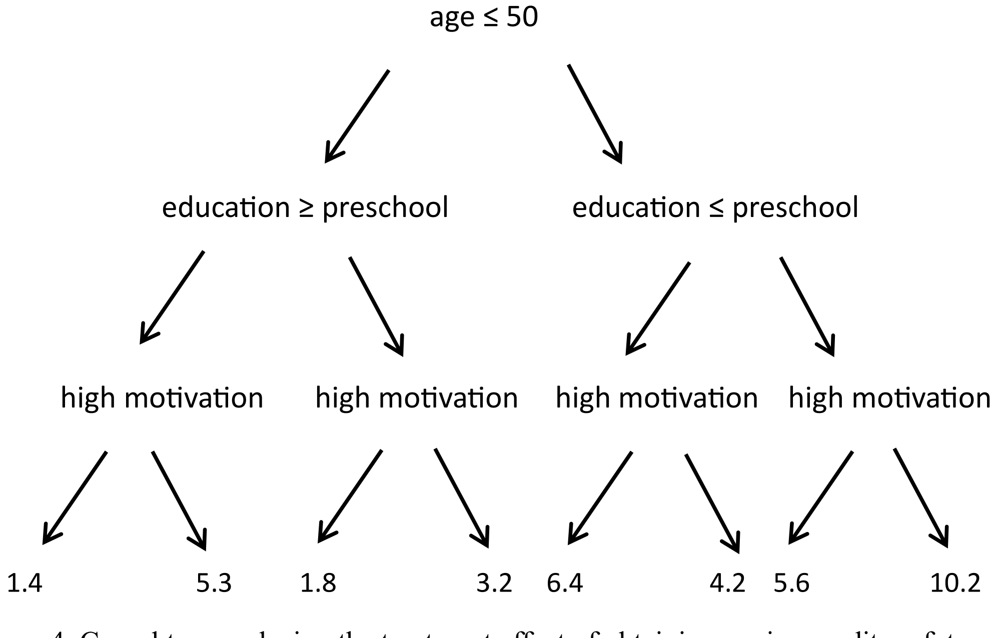

Figure 4. Causal tree analysing the treatment effect of obtaining a microcredit on future income.

Some machine learning proponents may not be overly concerned about *inconsistent variables* and may simply claim that the causal tree depicted in figure 4 could be extended so that it splits twice on age. This would solve the problematic issue of *inconsistent variables* for age in the provided example. Although it may be feasible to accomplish this for the presented case in figure 4, it is normally only possible to a very limited extent. The reason for this is that pruning and cross-validation techniques need to be applied to decrease the *sample variance* of the estimates as has been presented in chapter 2.1. Therefore, it is simply not possible to create a large number of splits since they would inevitably be partly removed through pruning and cross-validation strategies. Given that causal trees are built in settings in which many variables may impact the outcome, it is highly probable that many *inconsistent variables* are present in the control variable pool for generating the causal tree. This is the case as more variables can be employed in machine learning methods like causal trees than with standard econometric methods (Athey and Imbens 2019). Nevertheless, it is simply not possible to create that many splits as this would inevitably increase the *sample variance* of the causal tree algorithm. Consequently, taking the estimated subgroup treatment effects as an approximation for individual treatment effects becomes unreliable as the following example underlies:

Suppose a group of scientists is interested in examining the heterogenous treatment effects of a potentially life-saving drug. However, the drug has very strong side effects and not every patient shows a treatment response. Therefore, the scientists conduct an experiment to gain a better understanding of whom to give the drug. Nevertheless, there is still the potential issue of confounding since some characteristics may affect both the treatment and outcome variable. In order to address this issue, the scientists employ control variables as in the case of observational settings. As the researchers end up incorporating many different control variables, they conclude that machine learning methods in the form of causal trees are needed to analyse the heterogeneity of treatment effects. In addition, they do not have any prior knowledge of how to specify the respective subgroups. Therefore, the researchers do not want to establish the subgroups based on theory but prefer applying an algorithm for this task. As the scientists are aware of the overfitting risk, they apply pruning techniques which lead to the causal tree depicted in figure 5. After estimating the causal tree, the scientists assess the following result:

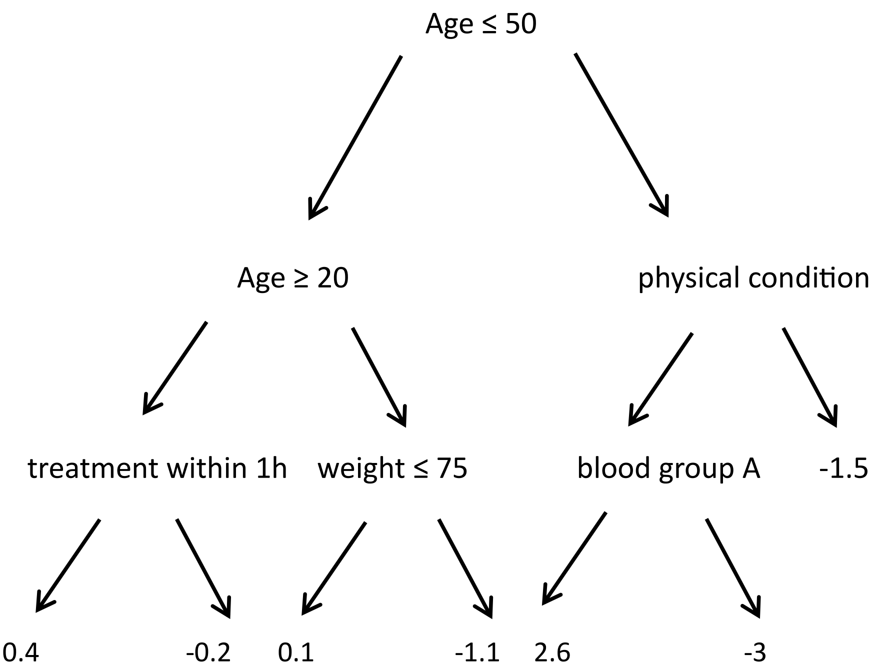

Figure 5. Causal tree with heterogenous treatment effects for a potentially lifesaving drug.

According to the causal tree, it seems that age, general physical condition, weight, blood group and the timing of the treatment are important drivers for different treatment responses. Now, even if one believes that the causal tree in figure 5 correctly analyses the situation of interest, the issue of *inconsistent variables* precludes the extraction of new knowledge for individual treatment effects from the causal tree. For example, age, weight and the timing of the treatment could be highly *inconsistent variables*, requiring many more splits to reliably determine individual treatment effect responses. Consequently, the causal tree would need to be partitioned into many more subgroups as they all entail different treatment effect responses. However, the knowledge of potentially *inconsistent variables* is hidden in the data and cannot be inferred from the causal tree with only 3 splits. Given the strong side effects of the drug, the subgroup splits are insufficiently reliable to take the estimated subgroup treatment effects as approximations for individual treatment effects. This is the case as the treatment effects possibly vary even within the subgroups to a large extent due to the presence of *inconsistent variables.* Therefore, the *individual variance* in causal trees with many *inconsistent variables* is high. Consequently, *inconsistent variables* necessitate more subgroup partitions in case the causal tree method is applied to estimate individual treatment effects. However, as has been pointed out before, splitting the tree into more subgroups leads to higher *sample variance* and hence, the risk for overfitting increases. In the case of the potentially lifesaving drug, even slight differences in treatment responses could save additional lives. However, causal trees do not seem reliable to gain this necessary knowledge without falling into the overfitting trap.

Importantly, the problem of *inconsistent variables* poses a greater challenge for causal trees than for conventional decision trees. While the individual predive performance of decision trees also deteriorates in the presence of *inconsistent variables*, the estimation of individual treatment effects is even more problematic. This is because of the honest estimation step, which was discussed in the chapter introducing causal trees. *Honest estimation* requires that half of the training set is used to estimate the treatment effect, while the other half is employed to identify the splitting variables (Wager and Athey 2018). As a result, fewer observations can be employed to generate the causal tree. Fewer observations entail that less information can be used to estimate the average treatment effects within the subgroups which in further consequence, increases the *sample variance* of the causal tree algorithm due to higher uncertainty (Hastie, Tibshirani, and Friedman 2009)[26]. In other words, causal trees have a high *sample variance* as the causal tree building process requires more data points than decision trees. Consequently, the *sample variance* of the causal tree in comparison to conventional decision trees is higher, making stronger pruning efforts necessary. Therefore, causal trees naturally split on less variables since pruning techniques will remove more splits than in the case of conventional decision trees. Hence, causal trees can incorporate less *inconsistent variables* due to the application of pruning and cross-validation techniques, entailing even higher treatment effect variation within the respective subgroups and thus, higher *individual variance*. Consequently, while the application of pruning and cross-validation techniques is needed to reduce the *sample variance* of causal trees, it can backfire for the estimation of individual treatment effects through increasing the *individual variance* in the presence of *inconsistent variables*. Given that causal trees employ many control variables, the problem of *inconsistent variables* is likely to occur when employing the causal tree method for individual treatment effect estimations.

As I have been demonstrating by the example of *inconsistent variables* in this chapter, the causal tree method is not reliable at estimating individual treatment effects whenever there exist several *inconsistent variables* that influence the treatment effect. While the inclusion of many *inconsistent variables* is needed to properly account for treatment heterogeneity, this is not possible due to the application of necessary pruning and cross-validation techniques, which aim to remove splits from the causal tree to reduce its *sample variance*. Consequently, I have argued that the causal tree method cannot properly account for *inconsistent variables*, entailing high *individual variance* as the information of the *inconsistent variables* cannot be included and

[26] The uncertainty is higher since less information can be gained about the phenomenon of interest due to fewer observations.

therefore, fails to correctly detect individual treatment effect heterogeneity. While this issue is also problematic in the case of pure prediction tasks and decision trees, it is even more severe for the causal tree method due to the application of *honest estimation*. As a result, I conclude that causal trees suffer from high *individual variance* characteristics and thus, taking the average treatment effect estimations as approximations for individual treatment effects, becomes unreliable.

4.2 Instability in causal trees

In addition to the previously analysed issue of *inconsistent variables*, causal trees also face high *individual variance* properties due to the instability of causal trees which will be discussed in this section. While there are various definitions of tree stability, this chapter focuses on the stability of causal trees as a grouping and clustering mechanism. Consequently, the following subchapter places less emphasis on the question whether causal trees consistently split on the same variables.

In other words, one is interested whether two causal trees constructed from the same sample data lead to comparable classifications of individuals. Achieving this goal does not necessarily entail obtaining the same causal tree as diverse causal trees can result in the same grouping results as I will now illustrate. Therefore, it is possible to establish the same subgroups through the application of different causal trees. For instance, imagine two distinct causal trees with only a single split creating respectively two branches. While the first causal tree splits on income, the second splits on education. Nevertheless, it can be the case that both causal trees yield identical results, meaning that the sample is divided into exactly the same two subgroups, despite splitting on different variables. Consequently, the grouping result is identical even though the causal trees differ. Because of this reason, the stability of causal trees in this chapter is solely examined in relation to causal trees as a clustering mechanism. Consequently, the splitting variables of the causal tree are not of primary concern.

The definition of stability employed in this chapter is closely related to Turney's (1995), who defines stability as follows: Stability "is the degree to which an algorithm generates repeatable results, given different batches of data from the same process. In mathematical terms, stability is the expected agreement between two models on a random sample of the original data, where agreement on a specific example means that both models assign it to the same class. The instability problem raises questions about the validity of a particular tree, provided as an output of a decision-tree algorithm. The users view the learning algorithm as an oracle. Obviously, it is difficult to trust an oracle that says something radically different each time you make a slight change in the data.” (25)

In contrast to most other machine learning proponents, Turney (1995) defined stability of decision trees in terms of class assignment. In other words, stable decision trees consistently assign observations to the same group if built from different batches of the same sample data. Hence, data from the same data generating process but different subsamples is used to build various decision trees and consequently, check decision tree stability. If the generated decision trees do not exhibit strong stability results, this is tantamount to the situation that individuals are put into different subgroups with different prediction results every time a different subsample is used to construct a decision tree. Consequently, if causal trees generated from the same data generating process group individuals into different subgroups with different treatment effect estimates, the *individual variance* of the causal tree is very high. In other words, individuals would be assigned to different subgroups with different treatment effect estimates depending on which part of the sample data is used to construct the causal tree. Hence, the estimated treatment effect for an individual could be different. This is the case as the generated subgroups differ in terms of the included individuals in case of high tree instability. Therefore, the *individual variance* of the treatment effect results is very high since individuals are assigned to different subgroups every time a causal tree is built from different parts of the sample data. As a result, every causal tree would indicate a different treatment effect estimate for the same individual, since the average subgroup treatment effect is taken as the approximate individual treatment effect. Consequently, this results in a high *individual variance*. Additionally, by construction the subgroup averages are maximally different from each other, meaning that being placed in a different group would often lead to a very different estimated treatment effect.

However, causal trees have never been examined in relation to their grouping stability even though they are often interpreted as a clustering mechanism primarily based on a weak interpretation as discussed before. Although no research has been carried out directly in relation to causal tree stability, I will show in the following section that insights from testing the grouping stability of decision trees can provide valuable intuition. This is because the causal tree methodology only entails small deviations from the decision tree algorithm as presented in chapter 2.3. Therefore, knowledge about the stability of causal trees can be inferred from analyses conducted on decision trees.

Despite the fact that only very few authors have been so far interested in the exact stability definition employed in this chapter, an analysis conducted by Jacobucci (2018) provides intriguing results. While Jacobucci (2018) focused on a broader definition of decision tree stability, he also assessed classification stability results by estimating the Jaccard coefficient. The Jaccard coefficient is a measure of similarity between two sets and ranges from 0 to 1. High values indicate that the two decision trees entail similar grouping performance, while low values imply a low tree stability (Hastie, Tibshirani, and Friedman 2009). For his analysis, Jacobucci (2018) estimated decision trees on 20 distinct real-world datasets. Based on this data, the author estimated 20 different decision trees on different subsamples of the data. To generalize the results, the employed datasets differed in their characteristics, entailing predictor numbers from 3 to 23 and different sample sizes. The results clearly indicate very low Jaccard coefficients, suggesting low result stability. Only one of the 20 datasets exhibited a stability coefficient higher than 0.5 with most values falling below 0.25. Consequently, it can be concluded that decision trees perform inconsistently as a clustering mechanism and thus, do not group individuals consistently into the same subgroups.

Given the described similarity between decision trees and causal trees in terms of their general mechanism, similar results can be expected when evaluating the stability of causal trees. The analysis conducted by Jacobucci (2018) is especially realistic due to the fact that real world datasets were employed. Additionally, the R code and datasets used in the study are publicly available, enabling easy replication. Thus, causal trees also need to be evaluated as an inconsistent clustering mechanism when it comes to establishing consistent subgroups based on different subsamples of the same data. Consequently, these results indicate that high *individual variance* results can be expected for causal trees. This arises from the fact that treatment effect estimations may differ for individuals as the causal tree algorithm leads to different grouping results, depending on the specific sample part of the data employed to estimate the causal tree. This means that individuals are classified into different subgroups with respectively different average treatment effect estimations since causal trees are likely to suffer from high grouping instability. Hence, it can be the case that different grouping results yield different individual treatment effect estimates for the same individual as the subgroup treatment effect is taken as an approximation for the individual treatment effect.

While some people may argue that alternative grouping indicators evaluating the stability performance of decision trees should be preferred instead of the Jaccard coefficient, the Jaccard coefficient values are of such low magnitude that employing alternative indicators appears very unlikely to produce divergent results. Furthermore, it is important to note that the stability coefficients were higher for smaller datasets with a low number of predictors (Jacobucci 2018). Considering that causal trees are supposed to be applied in high-dimensional settings with many control variables and a large sample size, it is likely that stability results are even worse in such scenarios. Therefore, it seems even more perplexing that causal tree proponents have so far neglected any performance checks of causal trees as a clustering mechanism. Although the classification stability results presented in this chapter cast doubt on the effectiveness of causal trees as a method to estimate individual treatment effects, there may be cases where the stability is high, even in a high-dimensional settings. However, further research is necessary to establish valid guidelines for interpreting causal trees in such cases. So far, the research conducted in relation to the stability of decision trees clearly indicates that causal trees must be evaluated as an unstable grouping mechanism entailing high *individual variance* properties and leading to unreliable individual treatment effect estimations.

4.3 Implications of high individual variance

As has been pointed out in the preceding two sections, causal trees are susceptible to high *individual variance* characteristics stemming from *inconsistent variables* and causal tree instability. In the following subchapter I aim to further elaborate on the implications of high *individual variance* for individual treatment effect estimation employing the causal tree method. As a measure of reliability for individual treatment effect estimation, I will use confidence intervals due to their intuitive interpretation. As has been elaborated in chapter 2.1, confidence intervals are important for reliable estimation results since they provide a measure of uncertainty and hence, reliability of the result. Moreover, there is a direct association between high *individual variance* and confidence intervals for individual treatment effect estimation. More specifically, causal trees with high *individual variance* properties lead to wide confidence intervals for individual causal tree treatment effect estimates. The relationship stems from the fact that causal tree estimates with high *individual variance* entail that the individual treatment effects within a specific subgroup significantly differ from the average treatment effect of the respective subgroup, thereby widening the confidence intervals.

As discussed in chapter 2.2, one of the aims of causal trees is to estimate individual treatment effects. Therefore, the sample data is split into different subgroups according to some characteristics (splitting variables). In further consequence, for every established subgroup, the average treatment effect is estimated. If now one wants to obtain the treatment effect for one individual outside the sample, the average treatment effect of the subgroup that shares similar characteristics with the individual is taken as an estimate for the individual treatment effect. However, as has been argued before, high *individual variance* properties undermine this strategy. This is the case as the treatment effect may highly differ between individuals in the same subgroup. To better illustrate this point, imagine the following example employing confidence intervals as a measure of reliability:

Suppose a policymaker wants to decide whom to provide with a microcredit. Therefore, she asks a group of researchers to conduct a study investigating the effect of microcredits on future income. The researchers decide to carry out a causal tree analysis and forward the results to the policymaker. For reasons of simplicity, assume that the causal tree only generated three different treatment effect subgroups with the following results depicted in table 1. As shown in table 1, three different subgroups based on age and the number of children have been created by the causal tree algorithm. Since the causal tree method faces high *individual variance* characteristics, the established confidence intervals for the estimated individual treatment effects are very wide as indicated by the values in the third column of table 1. As a result, the policymaker cannot use the results to decide whom to provide with a microcredit. Even though the employed causal tree established different treatment effect groups with different estimated treatment effects, the results for individual treatment effect estimates are unreliable. Because of that reason, it can equally well be the case that some members of the second subgroup (750€) have a higher treatment effect response than some members of the first subgroup (1000€). This is the case as both confidence intervals for estimated individual treatment effects are very wide. Because of that reason, the results cannot be employed for policy analysis as the uncertainty is too high to allocate microcredits to individuals based on the estimated confidence intervals.

Similarly, causal trees with high *individual variance* cannot be applied to provide doctors with individual drug recommendations since the variability of the drug´s effect is large within the established subgroups. In contrast, estimated individual treatment effect results with small confidence intervals provide policymaker with reliable results and only a small degree of variability. Hence, the example shows that the creation of various subgroups established by the causal tree algorithm does not have to provide researchers with good approximations of individual treatment effects. The reason for that is the high *individual variance*, which can be the result of *inconsistent variable*s or causal tree instability. Thus, the causal tree is unreliable for estimating individual treatment effects and fails to accomplish one of its main goals[27]. Consequently, this thesis is the first work to challenge the suitability of the causal tree method for the precise estimation of individual treatment effects.

Subgroup	Estimated treatment effect	Confidence interval
age ≥ 40, children ≥ 0	1000€	[125;1875]
age ≥ 40, children ≤ 0	750€	[0;1500]
age ≤ 40, children ≤ 0	850€	[0;1675]

Table 1. Estimated individual treatment effects with corresponding confidence intervals.

4.4 Biased results in causal tree estimates

The following subchapter aims to elaborate on the third challenge posed to causal trees in this thesis. Specifically, I want to introduce and discuss the problem of M-bias when applying causal trees, leading to the unreliable estimation of average treatment effects within subgroups. Biased results can occur when the method employed is systematically over- or underestimating the target value. In other words, the result is deviating from the true target value of interest hidden in the data. There are multiple reasons for these deviations. One of the possible reasons is often referred to as M-bias in the causal inference literature (Pearl 2000), which may affect the validity of the statistical results. As I will demonstrate in the following subchapter, M-bias is especially problematic in causal tree settings and thus, likely influences the results of the causal tree method. Consequently, it will be argued that the estimated average treatment effects within the subgroups may become unreliable as the causal tree method is highly susceptible to M-bias[28]. In order to facilitate the explanation of M-bias and its connection to causal trees, I will fall back on Pearl´s (2009) causal graph notation which was previously used to introduce the requirement of unconfoundedness in chapter 2.3.

4.4.1 M-bias in causal trees

Causal trees offer some advantages as the method allows to include more control variables compared to a conventional econometric framework[29]. Hence, researchers do not have to make

[27] Importantly, nothing has been said about the reliability of the causal tree algorithm to estimate subgroup treatment effects but only about the reliability to use the estimated subgroup treatment effects as approximations for individual treatment effects.

[28] While in the previous chapter it was shown that causal trees are unreliable at providing individual treatment effect estimates, this chapter also claims that the average treatment effect estimates within the subgroups become unreliable in the presence of M-bias.

[29] See chapter 2.4 for a detailed explanation of why that is the case.

decisions with respect to which variables they want to control for[30] but can control for every variable which potentially affects both the outcome and treatment variable. As was presented in previous chapters with the help of Pearl´s (2009) causal graph notation, controlling for variables is necessary to decrease the risk of confounding. In addition, it appears prima facie that controlling for many different variables in machine learning settings comes with little risk. Variables, that are not used by the causal tree method to split on, simply increase the reliability of the method as more variables are employed as control variables. In other words, controlling for numerous variables in the context of treatment effects lowers the probability of overlooking confounding variables that may open additional non-causal associations and lead to biased estimates. As has been already elaborated in chapter 2.3, the results are biased since not controlling for a confounding variable can result in spurious non-causal associations between the treatment and control variable.

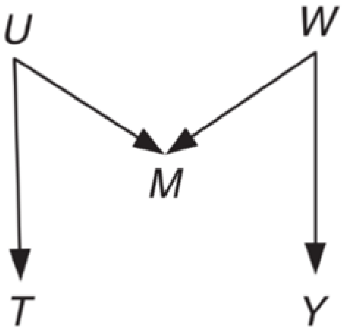

Figure 6. Causal graph M-structure (Ding and Miratrix 2015, 42).

Nevertheless, controlling for a large number of variables can also entail the opposite effect, namely introducing higher bias compared to controlling for only a few variables. This is the case since settings with many control variables are more susceptible to M-bias, which may arise when controlling for a potentially confounding variable (Pearl 2009). As illustrated in figure 6 M-bias is named according to its structure. While U and W are non-observable variables, M, T and Y are measured by the researcher. Like in chapter 2.3 which introduced the causal graph notation, one is interested in the effect of a treatment (T) on the outcome variable (Y). Without controlling for the observed variable M, figure 6 shows that T and Y are not causally connected (Ding and Miratrix 2015). However, when researchers believe in the importance of controlling for M[31], it opens a causal path between treatment and outcome variable, resulting in an

[30] As was discussed in chapter 2.4 controlling for too many variables in conventional econometrics settings inevitably leads to overfitting.

[31] This could be the case as M may appear to be an important confounding variable.

associative connection. This is the case as M is a common effect of U and W. Consequently, the causal path is now open once one controls for M. As described in chapter 2.3, a common effect can entail a non-causal association between two variables in case the common effect is employed as a control variable. Thus, stratified correlation arises between the unobserved variables U and W. In contrast, common cause variables, also called confounding variables, lead to non-causal associations by default. Their causal path can only be blocked by using them as control variables in the estimation process.

In figure 6, the parent variables U and W serve as the respective causes of T and Y. In the absence of additional measures, figure 6 shows that T and Y would not share any association as there is no direct causal arrow leading from T to Y. However, M is introduced as a control variable, which happens to be the common effect of T and Y. Consequently, a stratified correlation between U and W arises, as described in chapter 2.3. Importantly, U and W only share an association since M is included as a control variable in the estimation process. In further consequence, T and Y also share an association since they are connected to the unobserved variables U and W via causal arrows, as depicted in figure 6. Thus, controlling for M may lead to the erroneous belief of a causal connection between T and Y, resulting in biased estimation results (Pearl 2009). However, the non-causal association between T and Y is the result of M, which is the common effect of U and W. To gain a deeper understanding of M-bias in a treatment effect setting, one can consider the following example:

Suppose a researcher is interested in the effect of a microcredit on future incomes among individuals living in different villages. Since the researcher believes that the number of children (M) may causally affect future income (Y) and the effect of obtaining a microcredit (T), she decides to control for the variable number of children (M) to decrease the risk of confounding bias. The rationale behind that decision could be that the researcher believes that people having more children (M) are more likely to receive a microcredit, while the number of children (M) itself may also affect future income (Y) as having many children increases the work capacity. However, it is also plausible that factors causing variations in the number of microcredits obtained (T) and future income (Y) also impact the control variable number of children (M). For example, the specific village one lives in (U) may causally influence the number of children (M) people have. At the same time, living in certain villages (U) may causally affect the probability of receiving a microcredit (T) assuming that only people from certain villages can apply for it. Moreover, the age of a person (W) may have a causal impact on the number of children (M) and future income (Y). This could be the case because older individuals, on average, tend to have more children (M). Furthermore, there is a possibility that age (W) could causally influence future income (Y), assuming that younger individuals have more work years ahead of them.

Therefore, the number of children (M) is not a potential confounding variable requiring controlling for it but creates a causal M-structured graph as shown in figure 6. As a result, controlling for the number of children (M) leads to M-bias and hence, incorrectly suggests a causal link between the treatment variable (T) and future income (Y). In other words, an additional spurious correlation between obtaining a microcredit (T) and future income (Y) has been created through employing the number of children as a control variable in the causal tree building process. It is important to note that a M-structure can never be detected through statistical tests but can only be argued for from a theoretical perspective. As has been discussed in chapter 2.3, the same holds true for the issue of confounding. Consequently, had the researcher not presented theoretical justifications to control for the number of children (M), no association would have been identified between receiving a microcredit (T) and future income (Y). This demonstrates that controlling for many variables is not always advantageous to decrease the risk of biased results as it increases the probability for M-bias.

As revealed by the example given above, M-bias appears to be of special relevance for causal trees as this method controls for numerous variables. As a result, the likelihood for M-biases increases as every potentially confounding variable controlled for could give rise to a M-structure and thereby bias the outcome. Furthermore, given the inclusion of numerous variables in causal trees, scrutinizing these models for potential M-bias becomes an insurmountable task. As the precise challenges with M-bias and machine learning techniques like causal trees may seem abstract to some readers, the subsequent familiar example should provide an illustration of the issue at stake:

Suppose a researcher wants to group people into different heterogenous treatment effect subgroups after conducting a drug trial. Fortunately, a wide array of variables has been collected. Since the researcher wants to exclude for the possibility of confounding bias, she controls for every observed variable. Consequently, if the researcher wanted to check for the presence of M-bias, she would need to individually examine each control variable. First, she would have to provide sufficient theoretical justifications demonstrating how each variable may plausibly affect both the treatment and outcome variable, making controlling for the variable inevitably. Second, the researcher would have to examine every control variable for M-bias, presenting compelling theoretical arguments as to why it seems unlikely that two distinct variables causing the control variable independently affect the treatment and outcome variable, leading to a M-structured causal graph. In other words, researchers do not only have to provide theoretical justifications for the need to control for a specific variable, but also check every control variable for potential M-bias.

Hence, M-bias poses a particular challenge for the causal tree method due to two reasons. First, the probability of generating an M-bias increases as researchers control for a larger number of variables. Second, it becomes very time consuming to check in practice if a causal tree exhibits M-bias. This would necessitate examining each individual variable with the help of causal diagrams and theoretical evidence. While the fact that M-bias leads to biased results is widely accepted in the literature on causal inference, its prevalence is contentiously discussed. Economists like Rubin and Rosenbaum assert that M-bias is a rare phenomenon and thus, does not require much attention (Ding and Miratrix 2015). On the other hand, they state that the issue of confounding is more widespread in empirical research. Because of that reasoning, scientists should control for as many potential confounding variables as possible to decrease the risk for biased results in their research. Since M-bias is rare, researchers do not have to be overly concerned about generating a M-structure. In contrast, Pearl (2015) claims that M-bias is a frequent phenomenon and thus, requires scientists to be cautious about controlling for potential confounding variables. Consequently, M-bias cannot be neglected in empirical research settings. Even though I do not want to take a stance in the debate on the frequency of M-bias, it appears that it is a more prominent concern in the realm of machine learning methods. Standard causal inference settings start from controlling for a few variables to an approximated maximum of 15 variables (Angrist and Pischke 2009). In contrast, machine learning applications such as causal trees can control for a few hundred potentially confounding variables (Chernozhukov et al. 2017). Consequently, the risk for M-bias structures must be multiple times higher than in standard causal inference settings.

As has already been demonstrated in this chapter, both potential bias sources, M-bias and confounding bias, are theoretical concepts. Because of that reason, it is impossible to resolve the debate about the frequency of M-bias in empirical research. As emphasized in this chapter, M-bias can never be tested but only be argued for on a theoretical basis[32]. Consequently, variable relationships in causal graphs are frequently subject to intense debate. Therefore, M-

[32] As has been discussed in chapter 2.3, only correlative associations between variables can be tested. Thus, these correlative relations can only be given a causal interpretation with the help of theory or intuition.

bias frequency depends on the subjective standpoint of the respective researcher in how far the theoretical arguments for M-bias in specific situations appear credible. Nevertheless, it appears fair to conclude that the discussion from above revealed that M-bias is a bigger concern for machine learning methods like causal trees which control for numerous variables. Consequently, researchers applying causal trees face the insurmountable challenge of checking every control variable for M-bias. As this is rather time consuming, M-bias poses a problem for the causal tree method. In addition, as confounding bias and M-bias can only be argued for on a theoretical basis, the a priori standpoints of researchers play a role in detecting these biases, which in further consequence could influence the results of the causal tree method.

4.4.2 Simultaneous occurrence of M-bias and confounding bias

Assuming that causal tree practitioners engage in the time-consuming effort of examining every control variable for M-bias, another important methodological issue emerges. As argued by Pearl (2015), a critical concern arises when a variable holds the potential to introduce confounding bias if left uncontrolled, yet simultaneously leads to M-bias if controlled for. In other words, the results will be biased either way. The following section aims to provide some rough guidelines for the simultaneous occurrence of M-bias and confounding bias for causal tree practitioners. While it will be argued that researchers should be more concerned about confounding bias, the chapter demonstrates that the double bias problem cannot be resolved and thus, is likely leading to biased results when estimating treatment effects.

For machine learning methods like causal trees which control for many variables, the before described scenario may frequently be encountered by causal tree practitioners employing the causal tree method. In the literature often referred to as butterfly bias due to its structure, describes the complex situation of a simultaneous occurrence of M-bias and confounding bias (Ding and Miratrix 2015). As visualized in figure 7, the variable M is at the same time a potentially confounding variable and introducing M-bias when controlling for it. In other words, the butterfly bias is a situation in which there are good theoretical arguments that a variable introduces confounding bias if it is not controlled for. However, controlling for the respective variable creates a M-structure similar to the situation described in the previous subchapter. Therefore, researchers must decide which bias is potentially lower and consequently, would result in less biased outcomes.

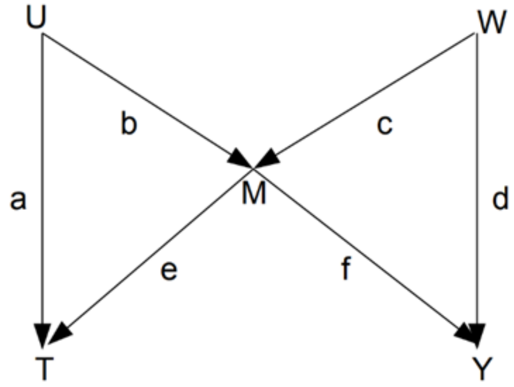

Figure 7. Simultaneous occurrence of M-bias and confounding bias (Ding and Miratrix 2015, 48).

In general, scientists should primarily pay attention to the theoretical arguments presented. Since both types of biases can only be argued for from a theoretical perspective, scientists can base their decision on which bias to accept by evaluating the credibility of the underlying theoretical assumptions. It is needless to say that in this case a solid theoretical argumentation is necessary to adequately justify the acceptance of one bias. Nevertheless, there may be situations in which both the theoretical reasoning for M-bias and confounding bias appears to entail similar credibility levels. Consequently, the discussion cannot be settled based on theoretical arguments. However, the properties of machine learning methods like causal trees may provide some practical advice for researchers.

As has been demonstrated by Pearl (2015), M-bias is weaker than confounding bias in noisy environments. In a nutshell, noise in a statistical environment can be understood as random variations in the data, obscuring the true underlying patterns in the data (Hastie, Tibshirani, and Friedman 2009). For example, inaccuracies in the measurement process of a variable can lead to statistical noise (Angrist and Pischke 2009). Since causal trees employ many variables in their analysis, it appears likely that some variables suffer from measurement errors and hence, lead to statistical noise in the causal tree. Statistical noise cannot be directly visualized in the M-structure shown in figure 7 but can be described as attenuating the associations between the parent variables (U, W) and the control variable (M). In other words, the estimated causal impact of U and W on M decreases as a consequence of statistical noise. According to Pearl (2015), the bias stemming from M-structure is only between 20,8-32,9% of the confounding bias in case both biases are present in a noisy environment. Consequently, researchers applying machine learning methods like causal trees should prioritize controlling for potential confounding variables over avoiding M-bias in cases both biases seem equally credible based on theoretical reasoning. This is the case as confounding bias is stronger than M-bias in settings with statistical noise. As has been described above, causal trees are likely to operate in high-noise situations since numerous control variables are used in the analysis, making measurement errors more likely.

Nevertheless, M-bias remains a significant concern for the application of causal trees since it leads to biased results, albeit with lower bias compared to confounding bias. In addition, it is impossible to determine the direction of the bias since the variables U and W are assumed to be unobserved (Pearl 2015). Their values would be necessary to calculate the correlation between U, W and M, to properly distinguish between a positive and negative bias in the causal tree results. Consequently, in the presence of both M-bias and confounding bias, accepting a M-structure due to its lower bias values gives rise to the issue of unknown bias direction, further complicating the interpretation of the causal tree method.

To conclude this subsection, I have argued that M-bias leads to unreliable results of the causal tree method mainly due to two reasons: First, causal trees include numerous control variables to decrease the risk of confounding, making it almost impossible for the researcher to check for potential M-bias structures. Second, when encountering double bias situations as a result of simultaneous M-bias and confounding bias, it is impossible to analyze the direction of the M-bias. Therefore, causal tree results become unreliable since researchers often do not know if M-bias is present in their research setting. In addition, even if they are aware of it, they do not have any idea in which direction M-bias influences the result.

4.5 Supplementary challenges for a strong interpretation of causal trees

While the previous chapter raised challenges for both possible interpretations of causal trees, this chapter aims to point out an additional difficulty when interpreting causal trees according to a strong interpretation. Consequently, I will demonstrate that upholding a strong interpretation of causal trees becomes more difficult, even if the before presented challenges have been addressed. In a nutshell, I show in this chapter that causal trees give rise to multiple different causal models and thus, cannot be given a strong causal interpretation. As has been done in previous chapters, I will mainly base my explanations on the causal graph notation introduced by Judea Pearl (2000) which facilitates explanation and visualization. While the main problem presented in this chapter is partly acknowledged by leading causal tree proponents (e.g. Athey and Imbens 2019), the consequences have never been thoroughly analyzed and scrutinized.

As discussed in section 2.3, causal trees establish heterogenous treatment effect subgroups without the need to prespecify the subgroups or splitting variables beforehand. In contrast, standard econometric methods require economic theory or intuition to determine the different treatment effect subgroups before estimating the econometric model. Despite this big advantage, causal trees cannot provide a deeper understanding of the factors underlying treatment heterogeneity. Hence, a strong interpretation is untenable. This is the case as one causal tree is compatible with multiple causal models. More specifically, causal trees do not necessarily split on the causal driving variables but can equally split on highly correlated counterparts without changing the grouping results (Chernozhukov et al. 2017). Consequently, if variables A and B are highly correlated and yield similar grouping outcomes, the causal tree algorithm is unable to differentiate between splitting on variable A or variable B. Consequently, in scenarios where only variable A truly acts as a causal driver for treatment differences, the causal tree algorithm may choose to split on variable B, with no substantive impact on the obtained results. Therefore, the splitting variables do not have to be the true causal drivers for the resulting causal tree but can only be highly correlated with the actual causal driver. To further strengthen this point, figure 8 and figure 9 show possible causal models which are both compatible with the causal tree in figure 4 (see p.33):

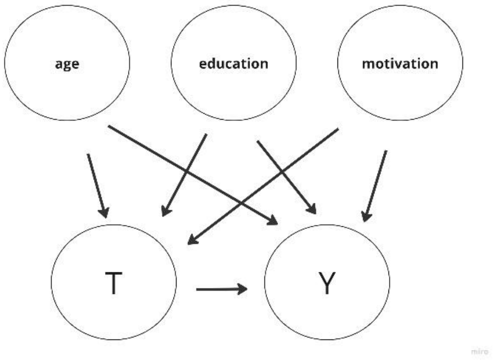

Figure 8. Potential causal model A

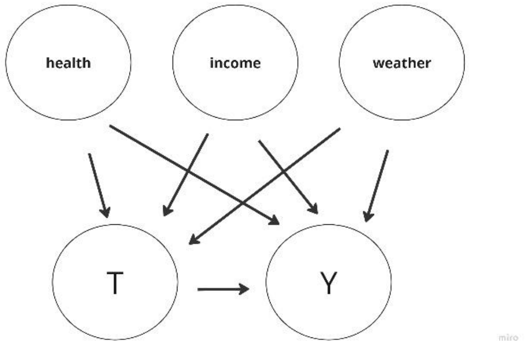

Figure 9. Potential causal model B

Both causal models could give rise to the same pattern of correlations, which in further consequence, lead to the same analysis by the causal tree algorithm. For example, there are strong empirical and intuitive reasons for a correlative connection between age and health. It seems uncontentious to make claims like: the general health condition tends to decline on average from a specific age onwards. Hence, there seems to be a correlation between health and age. Similar relations could be formulated for the other variables depicted in figure 8 and 9 (education/income, motivation/weather). Furthermore, it is important to note that both potential causal models A and B in the example entail that the unconfoundedness requirement is met since all causal back-door paths from the treatment to the outcome variable are blocked (Cinelli, Forney, and Pearl 2020). Given these correlative connections, the causal tree could have split on either of the connected variables as highly correlated variables yield the same grouping results (Chernozhukov et al. 2017). In other words, highly correlated variables may lead to the same grouping results, since the causal tree algorithm behaves similar in terms of its splitting process. This is the case as similar information, despite extracted from different variables, is employed in the causal tree building process. Consequently, a single resulting causal tree can encompass various causal models. However, it is impossible to detect the actual causal driving variables by employing the causal tree method, as there is no way to distinguish between potential causal model A and potential causal model B. Causal interpretation can only be added based on theoretical knowledge or intuition. While this point is partially recognized in the machine learning literature on causal inference, its implications are not extensively discussed (Athey, 2018).

To be more precise, if there are strong empirical or intuitive reasons to believe that a splitting variable is connected to another control variable, it is impossible to claim that the first variable causally impacts treatment heterogeneity. Given the typical high-dimensional setting with many variables in which the causal tree method is normally applied to satisfy the requirement of unconfoundedness, many variables may exhibit strong correlative connections. Therefore, several causal models could generate the same correlation pattern in the data and hence, generate the same causal tree in terms of grouping results. Consequently, the same grouping result can be achieved by different variable splits and thus, different causal trees.

Coming back to the two potential causal models A and B, their causal story differs in important aspects: while model A suggests that age, education and motivation are the driving forces for treatment heterogeneity, model B makes us believe that health, income and the weather are important causal factors for treatment differences. To draw reliable causal inferences, it is crucial to understand the causal relationships between the splitting variables and other potentially correlated variables. Theoretical knowledge is the only way to identify splitting variables as explanatory relevant. However, this entails an equally strong a priori theory commitment as in standard econometric modelling[33]. Consequently, researchers must possess knowledge of the relationships between variables before estimating the causal tree. As a result, the causal tree method loses one of its big advantages since the algorithmically driven splitting process alone is of little help for identifying the driving variables for treatment effect differences. Therefore, it appears that causal trees can only reliably infer causal relationships with similar theoretical commitments as those in a standard econometric setting.

However, causal trees were developed to avoid strong a priori theory commitments by not requiring the pre-specification of all potential subgroups before estimating the model (Athey, 2018). Therefore, it can be concluded that a strong interpretation of causal trees is unattainable as it is impossible to infer the causal drivers of treatment heterogeneity from causal trees. While the results of a causal tree can be visually represented in a causal tree diagram, it does not provide any insights into the underlying mechanism, as correlated variables with the selected splitting variables may be the actual drivers of treatment heterogeneity. Consequently, causal trees cannot be of any help when it comes to understanding the underlying mechanism of treatment effect heterogeneity. As I have argued in this chapter, this limitation arises because a single causal tree can give rise to various causal models due to the presence of highly correlated variables within machine learning settings. As a result, extracting information about the underlying mechanism responsible for treatment heterogeneity becomes impossible.

[33] It is important to remember that researchers need to prespecify the different treatment effect subgroups in a standard econometric framework through relying on theory or intuition.

5. Conclusion and outlook

In this thesis, I have analysed causal trees as a method for detecting heterogenous treatment effects, revealing essential limitations researchers should take into consideration when employing causal trees. My conclusions can be best summarized in the following three points:

Firstly, I have shown that causal trees fail to provide scientists with reliable individual treatment effect estimations, which constitutes one of the main goals of causal trees. Secondly, I argued that causal trees are incapable of reliably estimating average treatment effects within subgroups due to the complex interplay between M-bias and confounding bias, both of which are likely to occur when employing causal trees with many control variables. Thirdly, it has been established that causal trees cannot reveal anything about the underlying mechanism at work and thus, cannot help researchers to establish causal relationships. In order to arrive at the aforementioned conclusions, my argumentation was structured as follows:

In chapter 2, I introduced the concept of causal trees and elaborated on their advantages in comparison to standard econometric methods. Given that the causal tree algorithm is rooted in traditional decision tree methodology, my explanations were based on the latter. In addition, since causal trees have been developed for analysing treatment effect settings, the requirement of unconfoundedness was introduced and discussed with the help of causal graph notation. This served the following two purposes: enhancing the reader´s comprehension of the further argumentation provided in this thesis and allowing me to reveal the advantages of causal trees over standard econometric methods for detecting heterogenous treatment effects. First, I rehearsed the standard argument that causal trees allow for the inclusion of a larger number of control variables, thereby prima facie improving the credibility of results in relation to the unconfoundedness requirement. Second, I presented the argument that the causal tree method alleviates the a priori knowledge burden on researchers, as it algorithmically generates heterogeneous treatment effect subgroups. Moreover, chapter 2 illuminated a hitherto neglected trade-off when aggregating causal trees to a causal forest. Specifically, I argued that causal forest treatment effect estimations only possess valid confidence intervals when a limited number of control variables is employed. However, this undermines the advantage of causal trees to include many potential control variables to satisfy the requirement of unconfoundedness. Consequently, I offered a novel argument that causal forests cannot be considered a reliable method for the analysis of treatment effect settings.

In chapter 3, I introduced the distinction between a weak and strong interpretation of causal trees. Interpreting causal trees according to the strong perspective implies that the method can reveal the underlying mechanism, thereby disclosing the fundamental causes of heterogeneity in treatment effects. In contrast, from the perspective of the weak interpretation, the evaluation of causal trees focuses solely on the grouping results, disregarding the splitting variables or the understanding of the underlying mechanism. Consequently, employing a weak interpretation of the causal tree method implies that researchers cannot establish causal relationships with causal trees but only use it as a sophisticated clustering and prediction method. While a predominant number of authors advocate applying a weak interpretation, one can also find references in the literature endorsing a strong interpretation.

In chapter 4, I presented three challenges to the causal tree method, aiming to reveal its limitations. First, I introduced the novel concept of *inconsistent variables*, which are likely to be encountered in causal tree settings. As argued, *inconsistent variables* lead to high *individual variance* and therefore, undermine the ability of causal trees to reliably estimate individual treatment effects. Second, this thesis is the first work to examine the issue of grouping instability in causal trees. As I have claimed, causal trees are susceptible to suffer from high tree instability given the poor performance of decision trees as a clustering mechanism. Subsequently, I posited that this instability within causal trees increases the *individual variance* and thus, undermines the ability to estimate individual treatment effects. Third, I analysed the notion of M-bias within the context of causal trees and argued that the high number of control variables in causal trees makes it very difficult to detect M-bias. Consequently, causal tree estimation results are susceptible to bias. Furthermore, I discussed the increased possibility of simultaneously encountering confounding bias and M-bias in causal trees. Given that causal trees incorporate a large number of control variables, this situation holds particular significance for researchers employing this method. Moreover, I argued that causal tree proponents should rather tolerate the potential bias introduced by the M-structured causal graph instead of accepting confounding bias. Nevertheless, I claimed that causal trees should be regarded as an unreliable method for estimating average subgroup treatment effects given the increased likelihood of encountering either solely M-bias or the concurrent presence of both M-bias and confounding bias. In addition, in chapter 4, I also aimed to reject the possibility to interpret causal trees according to a strong interpretation, as one causal tree can encompass multiple causal models.

In lights of the arguments provided in this thesis, causal trees should not be promoted as a promising method to analyse treatment effect settings. Furthermore, future research should direct its attention towards the establishment of valid confidence intervals for causal forests with many variables in order to address the problem of high *individual variance*. Even though statisticians are sceptical regarding the feasibility of this endeavour, its realization would allow researchers to conduct precise individual treatment effect estimations, assuming no M-bias is present. Nonetheless, machine learning methods exhibit higher potential for transforming economic and social research practices within its original domain, which encompasses prediction tasks. For example, remarkable achievements in the field of macroeconomic forecasting have been documented in recent years. A notable example is the recently developed Macroeconomic Random Forest, a method combing standard linear regression with traditional machine learning methods (Goulet Coulombe 2020). Consequently, economists and social scientists should dedicate their time to modify and adapt machine learning methods in alignment with their specific research needs in relation to prediction tasks. Conversely, this thesis has shown that researchers should refrain from employing causal trees for the analysis of treatment effect settings.

6. Glossary

Active predictor variables: the variables that a decision tree or causal tree employs to generate different subgroups.

Algorithm: an automated mathematical procedure for making predictions and learning from the data.

Bagging: the practice of combining and averaging individual machine learning models to decrease the risk of overfitting.

Bias-variance trade-off: describes the need to balance bias and variance properties in statistical models.

Causal forest: an extension of the causal tree method that combines and averages multiple causal trees.

Causal trees: a machine learning method, originating from decision trees, modified for the analysis of treatment effect settings.

Cross-validation: a method for checking the generalizable capabilities of a model.

Decision tree: a machine learning method forming subgroups of the sample data with the primary aim of prediction and exploratory data analysis.

High-dimensional models: statistical models incorporating many variables with potentially non-linear relationships.

Honest estimation: a three-step process in the causal tree algorithm that separates the forming of subgroups and the estimation of treatment effects within these subgroups.

Individual variance: a measure of reliability of the estimated result when applied as a prediction for an individual.

Non-active predictor variables: the variables that a decision tree or causal tree does not employ to generate different subgroups despite being in the control variable set.

Overfitting: a phenomenon that occurs when a model captures too much noise or random fluctuations, leading to bad generalization properties.

Pruning: a method to remove splitting points that do not improve the predictive accuracy of a decision tree in order to reduce its variance.

Random forest: an extension of the decision tree method that combines and averages multiple decision trees.

Sample variance: a measure quantifying the uncertainty of a measurement or parameter estimation in a statistical model.

Splitting: the process of dividing the sample data into smaller subsets based on some characteristics/variables and a splitting criterion.

7. Bibliography

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. https://doi.org/10.1515/9781400829828.

Athey, Susan. 2018. 'The Impact of Machine Learning on Economics'. In *The Economics of Artificial Intelligence: An Agenda*, 507–47. University of Chicago Press.

Athey, Susan, and Guido Imbens. 2016. 'Recursive Partitioning for Heterogeneous Causal Effects'. *Proceedings of the National Academy of Sciences* 113 (27): 7353–60. https://doi.org/10.1073/pnas.1510489113.

Athey, Susan, and Guido W. Imbens. 2017. 'The State of Applied Econometrics: Causality and Policy Evaluation'. *Journal of Economic Perspectives* 31 (2): 3–32. https://doi.org/10.1257/jep.31.2.3.

———. 2019. 'Machine Learning Methods That Economists Should Know About'. *Annual Review of Economics* 11 (1): 685–725. https://doi.org/10.1146/annurev-economics-080217-053433.

Benini, Giacomo, and Stefan Sperlich. 2022. 'Modeling Heterogeneous Treatment Effects in the Presence of Endogeneity'. *Econometric Reviews* 41 (3): 359–72. https://doi.org/10.1080/07474938.2021.1927548.

Breiman, Leo. 2001. 'Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)'. *Statistical Science* 16 (3). https://doi.org/10.1214/ss/1009213726.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 2017. *Classification And Regression Trees*. 1st ed. Routledge. https://doi.org/10.1201/9781315139470.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018. 'Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India'. w24678. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w24678.

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2020. 'A Crash Course in Good and Bad Controls'. *Sociological Methods & Research*, 00491241221099552.

Davis, Jonathan MV, and Sara B Heller. 2017. 'Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs'. *American Economic Review* 107 (5): 546–50.

Ding, Peng, and Luke W. Miratrix. 2015. 'To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias'. *Journal of Causal Inference* 3 (1): 41–57. https://doi.org/10.1515/jci-2013-0021.

Fink, Günther, Margaret McConnell, and Sebastian Vollmer. 2014. 'Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures'. *Journal of Development Effectiveness* 6 (1): 44–57. https://doi.org/10.1080/19439342.2013.875054.

Gong, Xiajing, Meng Hu, Mahashweta Basu, and Liang Zhao. 2021. 'Heterogeneous Treatment Effect Analysis Based on Machine-learning Methodology'. *CPT: Pharmacometrics & Systems Pharmacology* 10 (11): 1433–43. https://doi.org/10.1002/psp4.12715.

Goulet Coulombe, Philippe. 2020. 'The Macroeconomy as a Random Forest'. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3633110.

Gulen, Huseyin, Candace Jens, and T Beau Page. 2021. 'The Heterogeneous Effects of Default on Investment: An Application of Causal Forest in Corporate Finance'. *Available at SSRN 3583685*.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer.

Hoover, Kevin D. 2013. 'The Role of Hypothesis Testing in the Molding of Econometric Models'. *Erasmus Journal for Philosophy and Economics* 6 (2): 43. https://doi.org/10.23941/ejpe.v6i2.133.

Imbens, Guido W. 2022. 'Causality in Econometrics: Choice vs Chance'. *Econometrica* 90 (6): 2541–66. https://doi.org/10.3982/ECTA21204.

Jacobucci, Ross. 2018. 'Decision Tree Stability and Its Effect on Interpretation.' Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/f2utw.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, eds. 2013. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. New York: Springer.

Keane, Michael. 2010. 'Structural vs. Atheoretic Approaches to Econometrics'. *Journal of Econometrics* 156 (1): 3–20.

Kent, David M, Ewout Steyerberg, and David Van Klaveren. 2018. 'Personalized Evidence Based Medicine: Predictive Approaches to Heterogeneous Treatment Effects'. *BMJ*, December, k4245. https://doi.org/10.1136/bmj.k4245.

Kotsiantis, Sotiris B. 2013. 'Decision Trees: A Recent Overview'. *Artificial Intelligence Review* 39: 261–83.

Kravitz, Richard L., Naihua Duan, and Joel Braslow. 2004. 'Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages'. *The Milbank Quarterly* 82 (4): 661–87. https://doi.org/10.1111/j.0887-378X.2004.00327.x.

Miller, Steve. 2020. 'Causal Forest Estimation of Heterogeneous and Time-Varying Environmental Policy Effects'. *Journal of Environmental Economics and Management* 103: 102337.

Pacifico, Antonio. 2021. 'Robust Open Bayesian Analysis: Overfitting, Model Uncertainty, and Endogeneity Issues in Multiple Regression Models'. *Econometric Reviews* 40 (2): 148–76. https://doi.org/10.1080/07474938.2020.1770996.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, U.K. ; New York: Cambridge University Press.

———. 2009. 'Myth, Confusion, and Science in Causal Analysis'.

———. 2015. 'Comment on Ding and Miratrix: "To Adjust or Not to Adjust?"' *Journal of Causal Inference* 3 (1): 59–60. https://doi.org/10.1515/jci-2015-0004.

Quinlan, J. R. 1986. 'Induction of Decision Trees'. *Machine Learning* 1 (1): 81–106. https://doi.org/10.1007/BF00116251.

Rokach, Lior, and Oded Maimon. 2014. *Data Mining with Decision Trees: Theory and Applications*. 2nd ed. Vol. 81. Series in Machine Perception and Artificial Intelligence. WORLD SCIENTIFIC. https://doi.org/10.1142/9097.

Stone, Charles J. 1982. 'Optimal Global Rates of Convergence for Nonparametric Regression'. *The Annals of Statistics*, 1040–53.

Wager, Stefan, and Susan Athey. 2018. 'Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests'. *Journal of the American Statistical Association* 113 (523): 1228–42. https://doi.org/10.1080/01621459.2017.1319839.