

Causal Forests

**How machine learning can enhance econometrics in uncovering
heterogeneous causal effects.**

Seminar New Models for the Digital Economy

Prof. Michael Burda, PhD

Humboldt-Universität zu Berlin
School of Business and Economics
Institute for Economic Theory II

by

Denis Augusto Pinto Maciel

(505471)

Berlin, July 20, 2019

Contents

1	Introduction	1
2	Causality & Experiments	2
2.1	The Oregon Health Insurance Experiment	2
2.2	Why experiments work?	3
3	Heterogeneity in Treatment Effects	6
4	Causal Forest	8
4.1	Decision Trees 101	8
4.2	Honest Causal Tree	10
4.2.1	Why causal?	10
4.2.2	Why honest?	11
4.2.3	From trees to the forest	11
5	OHIE & Causal Forests	13
5.1	Variable Selection	13
5.2	Modeling & Code	14
5.3	Results	14
5.4	Feature Importance	16
6	Final Remarks	18
A	Tables	21

1 Introduction

Economists are faced with two related, but very different type of questions. There is the question about prediction and there is the question about causality. This duality can be better illustrated by an example: the demand for hotel rooms. If we were given past data on room prices and occupancy of hotels, we would encounter a positive correlation between them: whenever the price is high, occupancy is expected to be high as well. Knowing nothing else, if you were asked to predict the price of a hotel whose occupancy is high, you should predict the price to be high as well. A naïve hotel owner, however, might conclude that by raising prices, he will increase the occupancy of his establishment. This, of course, doesn't make sense in the light of economic theory. The hotel owner is mixing up correlation with causation, prediction with causality. Demand for a good decreases as its price goes up. In the hotel case, there is a third factor that drives both demand and prices up. Vacation season increases demand for hotel rooms. Seeing this, hotel owners by raising prices. Thus, it becomes clear that the surge in demand due to high season is strong enough to make occupancy increase *despite* the simultaneous price increase.

The search for causality has been long one of the main concerns of economists. In recent years, machine learning - a field that has been mostly concerned with prediction - started to be used together with econometrics to answer questions about causal inference. Following these new developments, we will reassess in this paper the question about the causal effect of Medicaid coverage on visits to the emergency department. This problem has been already analyzed with battle-tested econometric methods by Taubman et al. 2014 . Here, however, we will provide an overview about causal inference, heterogeneity in causal effects and how decision trees can be stitched together to uncover causal effects. Finally, we will apply causal forests, a method whose basic structure was developed within machine learning (Breiman 2001), to try to identify heterogeneity in the response to Medicaid coverage.

2 Causality & Experiments

Does Medicaid¹ coverage increase or decrease the number of visits of a person to the emergency department? This question comes up whenever an expansion of Medicaid is considered. This is the case mostly because of the fiscal impact that an increase or decrease of emergency department visits resulting from more people having access to Medicaid could have. There are two lines of reasoning about what could be the effect of a Medicaid expansion on the state of public finance. The first one argues that by granting insurance to more people, one is to expect an overall higher usage of health care and thus an increase in public spending. A second line of reasoning defends that a Medicaid expansion could decrease insurance spending. Medicaid would allow participants to have greater access to primary care and by doing so it would be able to prevent acute conditions that result in an emergency department visit. Emergency department visits, it turns out, are much more costly than primary care. According to this line of reasoning, access to Medicaid and the ensuing increase in primary care (which is cheaper) would reduce the usage of emergency departments (which is considerably more expensive). This way, Medicaid would end up having a positive effect on public accounts.

2.1 The Oregon Health Insurance Experiment

In 2008, the state of Oregon opened a waiting list for the expansion of Medicaid. The program had been closed for new enrollments for over four years. As detailed in Finkelstein et al. 2012, the expansion targeted low-income adults that would otherwise not be eligible for Medicaid. More specifically, those eligible for the program were adults ages 19-64 who were Oregon residents, either U.S. citizens or legal immigrants, have not had health insurance for the past six months, had income below the federal poverty level and had assets below \$2,000. There were 10,000 available slots.

Oregon has correctly anticipated the high demand for the program and decided to run a lottery to pick the winners. This is the perfect scenario for identifying and disentangling causal effects of Medicaid. As explained below, the randomization in treatment assignment coming from the lottery draws enable us to isolate the effect of treatment from other otherwise correlated covariates.

Registration for the lottery was available for one month from January 28 to February 29,

¹Although frequently mentioned together in the media, Medicaid and Medicare are two separate programs. Medicare is a federal program that provides health coverage for adults ages 65+ and disabled people irrespective of their incomes. Medicaid is both state and federal and provides health coverage for low-income adults.

2008. The state went on to publicize the lottery and the barriers to sign-up were relatively low: anyone could sign-up by telephone, by fax, by mail, in person or online. At the end of the five-week window, 89,824 individuals were registered in the lottery list. During the following months (March through September 2008), the state conducted eight lottery drawings. In each one of them, roughly the same number of participants were selected. As soon as an individual was selected, all members of her household became eligible for entering Medicaid. 35,169 individuals from 29,664 different households were selected by lottery. About 30% of the selected individuals ended up successfully enrolling in the program. Only 60% of the selected individuals sent an application and from the applications sent, about half of them were deemed ineligible.

The data used in our paper was made available by Finkelstein et al. 2012 and can be downloaded at <https://www.nber.org/oregon/4.data.html>.

2.2 Why experiments work?

In an ideal world (with observable parallel universes), the causal effect of Medicaid coverage on emergency department visits could be estimated for every single individual by subtracting the number of visits under Medicaid from the number of visits without Medicaid that he or she had. For obvious reasons, a person can either be under treatment or control. We can never observe both states of the world simultaneously. The impossibility to observe the same person under treatment and control has come to be known as the *fundamental problem of causal inference*.

Instead of individual treatment effects, an alternative is to try to estimate the *average* treatment effect of a population (Morgan and Winship 2015). To make what follows understandable, it's important to first define some basic notation:

- Y_0 : dependent variable Y when under control
- Y_1 : dependent variable Y when under treatment
- $T \in \{0, 1\}$: dummy variable indicating whether observation is under treatment or control
- π : share of population under treatment

With that in mind, the average treatment effect can be defined as:

$$ATE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \tag{1}$$

By taking into account the actual treatment taken by each share of the population, we can rewrite the equation above as:

$$ATE = \{\pi\mathbb{E}[Y_1|T = 1] + (1 - \pi)\mathbb{E}[Y_1|T = 0]\} - \{\pi\mathbb{E}[Y_0|T = 1] + (1 - \pi)\mathbb{E}[Y_0|T = 0]\} \quad (2)$$

Formulating the problem in terms of average treatment effects does not by itself free us from the fact that some states are unobservable. In equation 2, the expected value of Y_1 for the share of population under control ($\mathbb{E}[Y_1|T = 0]$) and the expected value of Y_0 for the share of population under treatment ($\mathbb{E}[Y_0|T = 1]$) cannot be observed. In general, groups under treatment and control are different and have, for this reason, different expected values of the dependent variable. For example, in the absence of an experiment setting, a group of people taking a new drug with unknown side-effects (treatment) is in all likelihood very different from those not taking it (control). Those taking the drug could be much sicker than those not taking it and, therefore, willing to take higher risks to achieve a potential cure. In this scenario, comparing the average health (the dependent variable) of treatment and control wouldn't make sense, since those in treatment have already worse health and, even if the drug has a beneficial effect, treatment's average health would still be lower than control's.

This, however, wouldn't be a problem if we knew exactly which variables are responsible for the differences in treatment and control. We could simply control for them (i.e., include them in the regression as dependent variables) and our estimates of treatment effects would be accurate. Yet not only can we not be sure about all variables that drive differences in treatment and control groups, but also some of these variables are not observable in the first place. Going back to Medicaid and emergency department case, it's plausible that the psychological traits of individuals do play a role in the number of visits to the emergency department they have. A person can have many visits to the emergency department either because she is truly sick or because she has a more cautious personality. Although her health condition is observable (we can measure her blood pressure and glucose levels, etc), the psychological condition of someone being overly cautious and visiting the emergency department at the first suspicion of illness is not. Unobservable variables cannot, per definition, be controlled for making it impossible in an observational study to disentangle its effects from the variable we are trying to analyze. In such situations trying to estimate the effect of Medicaid on emergency department visits will probably be affected by omitted variable bias. Specifically, since we expect overly cautious people to be already in Medicaid, we would estimate that the effect of Medicaid on increasing emergency department visits is higher than what it is in fact.

Random assignment of treatment frees us from the two problems mentioned above. By randomly selecting participants to receive the treatment, we expect treatment and control groups to be equal in every single aspect except for the treatment variable. Thus, any difference arising between the two groups must be caused by the variable whose causal effect we are trying to estimate. On the one hand, we can estimate average treatment effects directly by computing $ATE = \mathbb{E}[Y_1|T = 1] - \mathbb{E}[Y_0|T = 0]$, since $\mathbb{E}[Y_1|T = 1] = \mathbb{E}[Y_1|T = 0]$ and $\mathbb{E}[Y_0|T = 1] = \mathbb{E}[Y_0|T = 0]$. On the other hand, random assignment breaks the correlations between unobservable variables and our treatment, because the latter was randomly assigned.

The subjects of the Oregon Health Insurance Experiment (OHIE) were randomly assigned to the treatment (Medicaid). This is a perfect setting for assessing the causal effects that Medicaid might have in other variables. The first step in such an analysis is to make sure that the experiment was able to balance control and treatment groups. For that, we compare both groups with respect to the variables we can observe. In OHIE, we compared mean year of birth, percentage of female, percentage that provided a phone number, percentage that signed up for lottery list on the first day across treatment and control. Table 1 shows that the difference between treatment and control across all metrics are very small and none is significantly different from zero. This is indicative that the lottery has managed to create comparable groups and that any significant variation in emergency department visits can be attributed to access to Medicaid or lack thereof.

Metric	Control Average	Treatment Average	Difference
year_of_birth	1968.34	1968.50	-0.16 (0.21)
pp_female (%)	55	54	1.1 (0.7)
gave_phone_number (%)	87	88	-1.0 (0.8)
first_day_list (%)	9	10	1.0 (0.7)

Table 1: Comparison between groups under treatment and control

3 Heterogeneity in Treatment Effects

Heterogeneous treatment effect is a mouthful, but ultimately a very simple concept. It refers to the fact that different people might react differently to the same treatment. Take the fictional example of a \$1000 transfer to individuals from a specific population. This is our treatment. Imagine there are two groups in our population: drug addicts and non-drug addicts. We want to measure the impact that this transfer has on the health of the individuals. Here we measure health with an index that ranges from 0 to 100. In this fictional example, \$1000 cause the health of drug addicts to decrease because they will use the money to acquire more drugs. For non-drug addicts, the effect on health is positive.

Table 2: Simple Heterogeneity Example

	\$1000	\$0	Treatment Effect
Non drug addicts	64	62	2
Drug addicts	17	25	-8

If the population is comprised of 95% of non-drug addicts, the estimate of the average causal effect (ATE) of a \$1000 transfer on health would be of 1.5 points in the health index.

$$ATE = 2 \times 0.95 + (-8) \times 0.05 = 1.5 \quad (3)$$

Across the whole population, the average treatment effect is positive, even though for the drug addict group the treatment has a nefarious effect. A well-meaning policymaker would probably want to identify these two groups and target them differently.

It's easy to see how heterogeneity can take much more complex forms. We need not stop in differences across two groups, we can think as many groups as there are characteristics identifying the units of observation. In the specific case, one could think of age, gender, and marital status as also influencing the response of the individuals to treatment.

Researchers haven't been blinded to heterogeneity in causal effects. On the contrary, many methods, very complex and otherwise, have been developed to disentangle different groups with different reactions to treatment (Imbens and Angrist 1994). Usually, the identification and estimation of treatment effects take the form of a specification plan done before the data collection where the researcher determines which covariates she suspects may be a source of heterogeneity. The pre-specification aims to refrain researchers from cherry-picking covariates after data collection that seems a statistically significant source of heterogeneity. After all, if

enough covariates are measured, some combination of them will by chance pass the statistical tests for heterogeneity.

With the rise of big data, however, researchers have been faced with yet a different problem. On the one hand, they want to make use of every piece of information that is relevant for identifying heterogeneous treatment effects. But there is the problem that if one looks long enough, she will find heterogeneity that comes to be due to sampling. The tension between the scientificity of a method and the desire to extract relevant information from the data can be softened with the aid of machine learning.

4 Causal Forest

4.1 Decision Trees 101

Imagine you are in charge of the lending department of a bank. You receive daily many requests of people wanting to borrow money from your employer. Along with those requests, you have access to a myriad of information on each applicant: age, gender, income statement, marital status and so on. You also have the same data on people who have gotten loans in the past. For the past loans, you also know if they have been paid back or the borrower has defaulted. Your task is to decide which of the new applicants should be granted a loan and you should pick ideally only those who won't default should.

One way of doing this is to apply a decision tree. In its most general form, a decision tree is a set of rules that splits a population into smaller groups. In our example, we want to split our applicants into subgroups and for each one of them, we either grant or reject a loan to the applicants belonging to it. The way you come up with rules can be arbitrary. You could probably use common sense and only grant loans to those with income over a certain threshold. In this case, you would end up with a very shallow tree with two leaves (also called end nodes): one leaf would contain those who make more than your threshold and the other one would contain those who make less. Nothing is limiting the number of nodes in your tree. In fact, you could grow your tree to have as many nodes as there are applicants (as long as the information you have on them is enough to uniquely identify each one of them). For example, you could be more restrictive with loan granting and after the split based on income, you further split the node of high-income participants into younger and older than 25. Only those applicants with high income and above 25 years would be granted a loan.

Although decision trees can accommodate a variety of heuristics when building the tree, machine learning practitioners take a more principled approach when deciding which rules to pick in order to split each node. Specifically, they look at past data and try to find rules that nicely splits the dataset according to a metric.² Here, we would look at data on past loans and try to minimize the number of defaults, while still granting loans to solvent applicants. To make matters simple, let's assume that to grant a loan to a defaulting applicant is as costly as rejecting a loan to a creditworthy applicant (i.e. a false positive is equivalent to

²Many metrics can be used to quantify prediction error. In regression, the root mean squared error (RMSE), defined by $\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$, is normally used. In classification, there are various metrics to be chosen from. The choice of a metric is dependent on the nature of the problem. Accuracy (the ratio between the number of correctly predicted classes and the total number of observations) would be a poor choice in the case the task is to identify individuals with a rare disease.

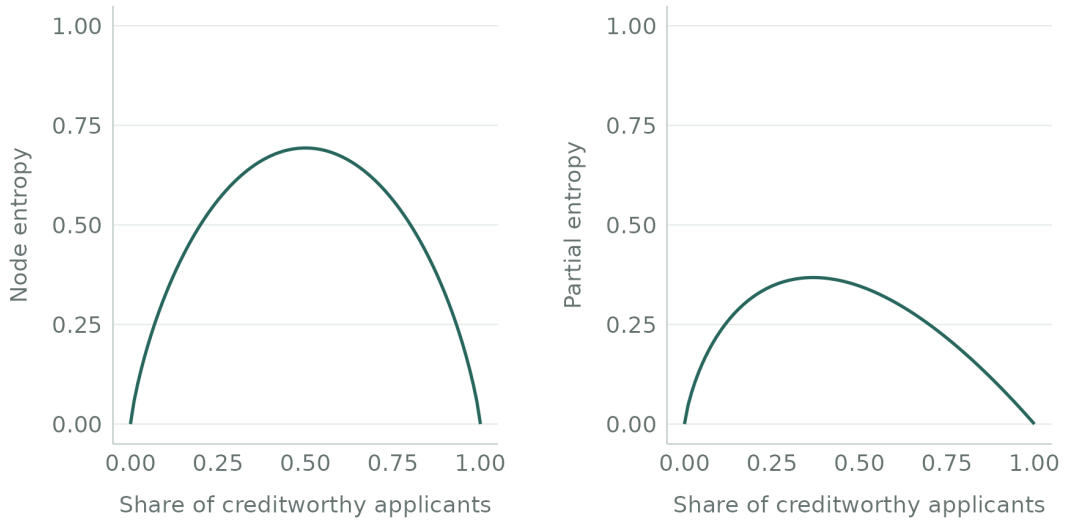


Figure 1: Entropy

a false negative in terms of cost). With this assumption in mind, we can use as our metric the accuracy, the ratio between the number of correctly identified applicants and the total number of applicants.

Finally, we need a loss function. Something that our algorithm will try to minimize as the ultimate criteria when picking splitting rules. For classification tasks, it's usual to pick entropy. In classification with n classes, the entropy H for every node S in a tree is defined as

$$H(S) = \sum_{i=1}^n -p_i \ln p_i \quad (4)$$

where p_i is the proportion of class i present in node S .

In our two-class example, the entropy in each leaf boils down to $H(S) = -p_c \ln p_c - p_d \ln p_d$. Figure 1 illustrates how entropy in a node varies in a two-class scenario with the proportion of one of the classes. Node entropy is the total entropy in the node, while partial entropy shows only the contribution of one of the classes to total entropy. To minimize entropy, our algorithm must create splits with as many applicants of one type as possible. The entropy is minimized when the share of creditworthy applicants is either zero or one. In other words, the entropy at a certain leaf will be zero if we manage to find a rule that leaves this leaf with either only creditworthy applicants or only defaulting applicants.

A popular splitting algorithm is what is called greedy search (Grus 2015). It goes as

follows: starting from the entire sample as you first node...³

- If the data in the node have all the same label (either creditworthy or defaulting), stop.
- Otherwise, try partitioning the applicants by each of the attributes (age, gender, etc).
- Choose the splitting rule that results in two new nodes with the lowest entropy weighted by the number of applicants in each node.
- Recur on each newly created node.

Once your tree is grown – that is, you have a set of sequential rules that puts each observation in a node –, you now can use it for prediction. You just apply the set of rules to an unknown applicant. If she ends up in a node with a majority of defaulting, you classify her as a default. If the node, however, is composed mainly of creditworthy applicants, you grant her a loan.

4.2 Honest Causal Tree

It turns out that slightly adjusting the decision tree can make it a powerful tool for identifying heterogeneous causal effects. An honest causal tree is a modified version of a decision tree that comes with the search for causal effects baked in (Wager and Athey 2018, Athey, J. Tibshirani, and Wager 2019).

4.2.1 Why causal?

We call it a causal tree because the tree tries not to minimize the entropy, but rather to maximize the causal effect in each node. As seen in the previous section, a typical decision tree splits the data by finding which pair of variable and value (for example, the variable could be age and the value 25) can most reduce entropy in the nodes if we were to split the node at it. Since we are interested here in treatment effects, we need to adjust this reasoning a little. First, we cannot observe the same individual under the two states (treatment and control). The best we can do is to identify individuals that are similar and compute the average treatment effect among them. In that vein, we might assume that the treatment effect for all individuals belonging to the same leaf of a tree is the same and can be approximated by the difference between the average of those in treatment and the average of those in control

³This simple version of greedy search could also be complemented with stopping rules related to the depth of the tree and the number of observations in each node. An example of the first case would be: stop when a branch of the tree is made by more than 5 rules, stop. For the second case: stop if the node has less than ten observations in it.

within that leaf. Thus, while building the tree, we pick the splits (combinations of variable and value) in such a way that the average treatment effect in the resulting nodes is maximized. In this sense are those trees causal: they have the search for causal effects baked in at their very core.

4.2.2 Why honest?

There is another peculiarity of the algorithm used to build causal trees worth mentioning and it's also the reason why they are called honest. The data used to 1) construct the set of rules that specifies the nodes and to 2) estimate the treatment effect of each node are different. In fact, the first step of the causal tree algorithm is to split the training data into two: a *splitting* sample and an *estimation* sample. The tree is then grown on the splitting sample only without ever having access to the estimation sample. Once the tree is fully grown, it's now time to estimate the treatment effect for every leaf in the tree. For that, the estimation sample comes into play and all observations in it are passed through the tree ending up in their respective leaf. Finally, the predicted treatment effect of each leaf of the tree is set to be the difference between the average of individuals in treatment and the average of individuals in control across all the observations (remember, coming from the estimation sample only) that ended up in that leaf.

4.2.3 From trees to the forest

As mentioned before, decision trees are very flexible and can accommodate an arbitrarily large amount of rules. Remember: decision trees can be as deep as having one node for each observation in the training set. You wouldn't expect such a tree to generalize to new data. As a nonparametric method, it will likely lead to overfitting (as in the case of one node per observation) unless it is regularized. For decision trees, regularization usually takes the form of bagging (Breiman 1996): multiple subsamples from the data are created, the model being regularized is trained on them, and the resulting multiple models have their predictions averaged out into one single model.

A random forest is an example of bagging methods. It is a collection of multiple decision trees that together deliver a single prediction. Each tree of a random forest is trained in a slightly different version of the data. Trees are constructed on random subsamples of both observations and features. The number of observations and the number of features that go in each subsample is the model's hyperparameters, i.e. they must be chosen by the researcher before the random forest is trained. After the random forest is trained and it's

supposed to deliver predictions on new data, each tree of the random forest predicts on the new observations and the average of those predictions is the general prediction of the random forest.

This procedure has the benefit of preventing overfitting, a technical term that means your model won't generalize to new data because it is fitting noise (randomness) in the training data. Most importantly, however, for causal forests (a random forest built out of causal trees), Athey, J. Tibshirani, and Wager 2019 have derived that the estimated treatment effects are asymptotically normal. This means that when your sample size gets larger, the distribution of the treatment effect estimates for every end node of our tree approximates the normal distribution. It's then possible to build confidence intervals around those estimates and assess the inherent uncertainty stemming from them.

5 OHIE & Causal Forests

Having solidified our understanding of the mechanics of causal forests, we are now ready to apply them to the Oregon Health Insurance Experiment (OHIE) data. Taubman et al. 2014 have estimated the average treatment effect of Medicaid on emergency department visits to be 0.408 with a p-value lower than 0.001. This result is, therefore, statistically significant and it means that a person under Medicaid is expected to have 0.408 more visits to the emergency department than what she would have had with no insurance. We reproduce this result in the analysis code under <https://github.com/denismaciel/causal-forest-ohie/tree/master/R>. With causal forest, however, we will output one treatment effect estimate per person and by doing that, try to capture the eventual heterogeneity in the data.

5.1 Variable Selection

The datasets of OHIE combined have more than 200 variables. This is a high number of dimensions to analyze and is probably well beyond any researcher’s ability to consider every possible interaction between those variables⁴ and how these interactions might affect the number of visits a person would have with and without Medicaid.

Here we will consider only 40 out of the over 200 different variables. An exhaustive list of the variables used to train our causal forest can be found in Table 4 in the Appendix. We have removed all variables that were collected *after* the lottery. The reason for that is that policymaking is one of the relevant areas causal forests can potentially be applied to. To include information that only became available after the treatment has been applied would prevent the results from being used in many policy applications, since the policy decision itself is often about to whom treatment should be applied and (for obvious reasons) one doesn’t have access to data collected after the application of the treatment at the time such a decision is made.

We have also excluded variables whose meaning not clear from the data dictionaries provided by the authors of the original study. For example, there are some variables related to previous participation in other welfare programs offered by the state of Orgeon. Although these variables might as well improve our causal models, it adds a bit of overhead to the reader and to the analysis itself. For that, we opted to leave them out.

⁴If we allow for a causal tree to be constructed with only 2 variables, there 40,000(200×200) different trees one could build

5.2 Modeling & Code

Before modeling, we have randomly split the data into a train and a test set. The train set contained 80% of the observations (19,674) while the test set contained the remaining 20% (4,918). We wanted to have an as close as possible overview of how the model would predict the causal effect for new, completely unseen data. Even though the splitting-estimation-sample procedure, which is already embedded in the causal tree algorithm, ensures that not the same data is used to grow the tree and to estimate the causal effects in the nodes, we go a step further and set aside 20% of the data, which will remain completely unseen by our causal forest. By doing so, we hope to have a better understanding of how the model will perform on out-of-sample data, which, again, is very important for public policy.

For the modeling itself, we used the `grf` R package⁵, which is maintained by the authors of the Generalized Random Forest paper. We have used the `instrumental_forest`, which is a generalization of `causal_forest` that accept an instrument together with the independent variables, the treatment and the dependent variable. The causal forest was composed of 2,000 trees. This is the number of trees recommended by the authors if you wish to also estimate confidence intervals along with the treatment effects.

5.3 Results

After training the causal forest, we use it to predict the causal effect of Medicaid on the number of visits to the emergency department on the observations in the test set. Figure 2 shows the estimated causal effects for every observation in the test set ranked by the size.

The estimated individual causal effects range from -4.8 to 4.9 with two extreme outliers at -9.8 and 5.4. Compared to the average treatment effect estimate of the original paper (0.408), it seems to indicate that there might be in fact heterogeneity in the effects and that enrollment in Medicaid could increase emergency department visits for some people and decrease it for other people.

Here we are visualizing almost 5 thousand data points and it's difficult to identify where the majority of the observations concentrate: are there more people with negative or positive treatment effects? Figure 3 answer this question by making clear that most of the predictions concentrate on the right side of the line of real numbers. This is in line with the original findings and is indicative that Medicaid leads to an increase in emergency department usage for most people.

As mentioned, one of the greatest features of causal forests is not only to provide estimates

⁵The package is open-source and can be found at <https://github.com/grf-labs/grf>.

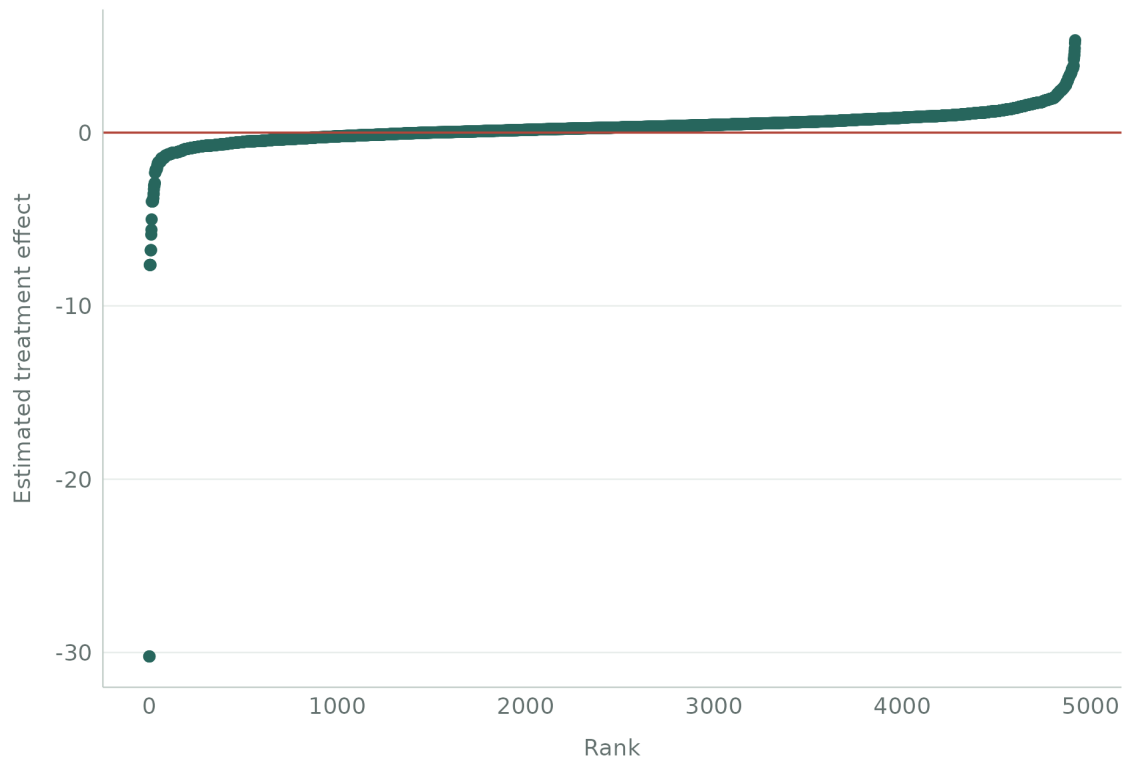


Figure 2: Ranked Treatment Effects

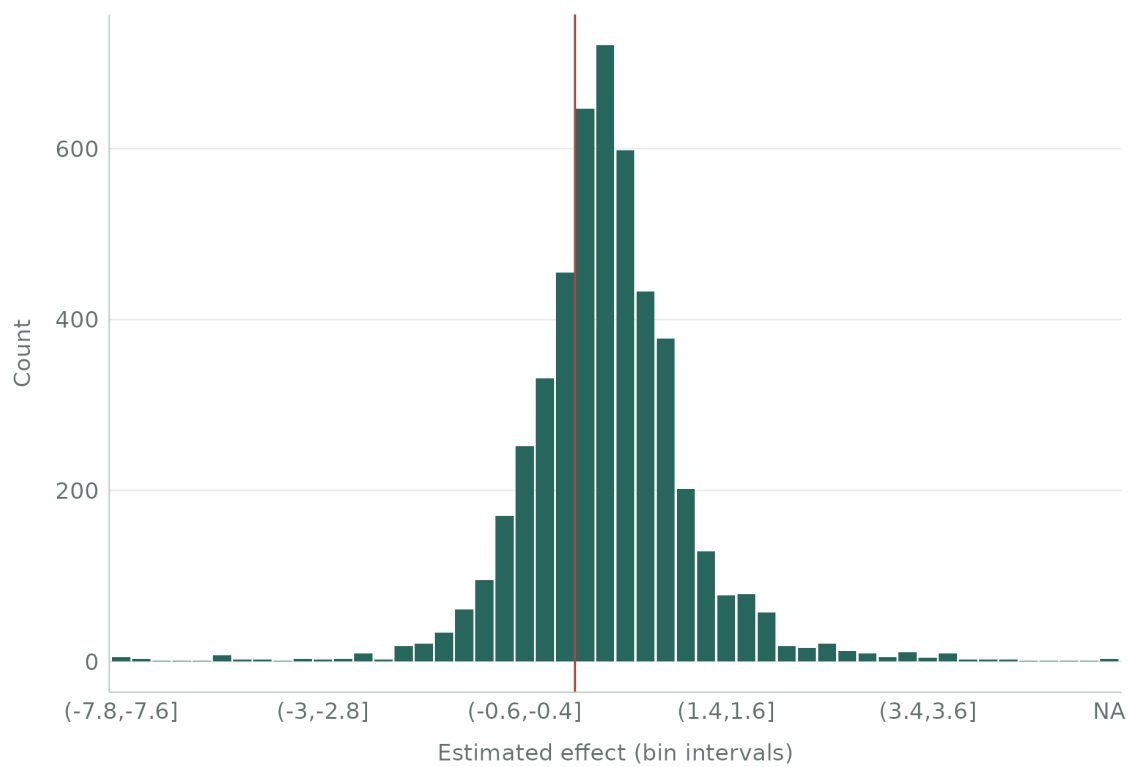


Figure 3: Treatment Effect Histogram

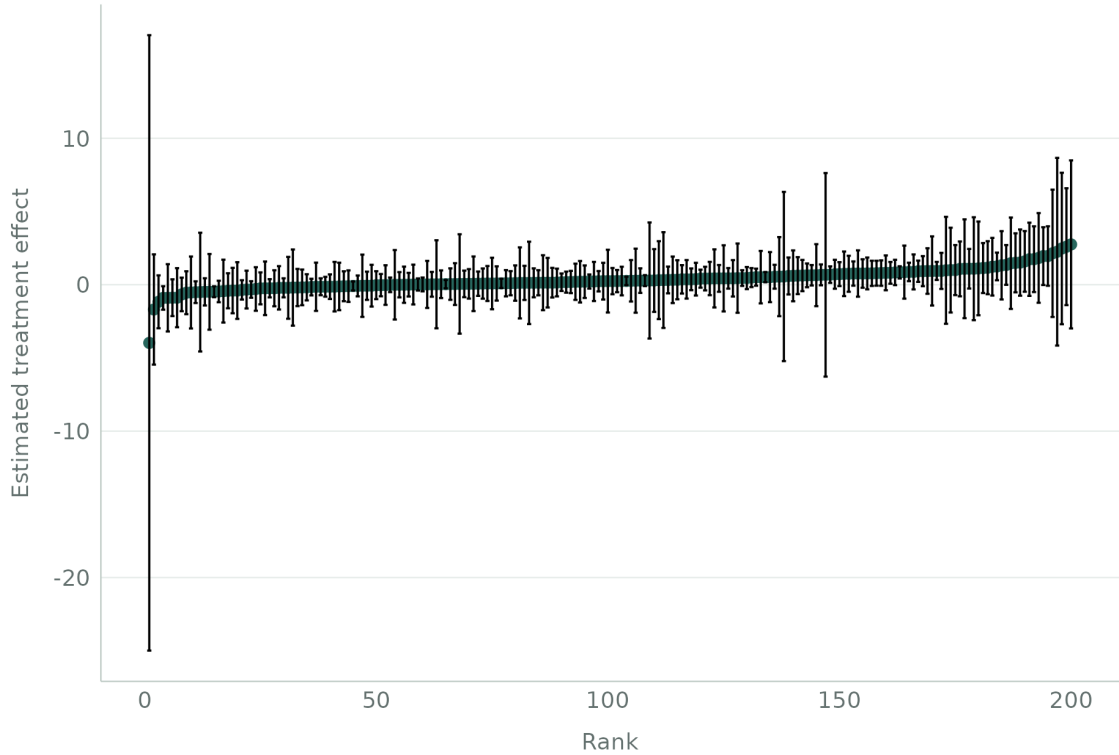


Figure 4: Estimated Treatment Effects & Confidence Intervals

on the heterogeneity of treatment effect but also that these estimates are asymptotically normally distributed. Armed with that, we can construct confidence intervals around our point estimates. In Figure 4, we have sampled 200 observations from the test set and displayed the predicted treatment effect estimates accompanied by their respective confidence intervals. There is a great amount of uncertainty on the estimates: even for the highest positive estimates, we cannot reject the null hypothesis with 95% confidence that they are greater than zero.

5.4 Feature Importance

Although many machine learning models are known for being difficult to interpret (black box models),⁶ feature importance is an interesting metric that allows us to identify which variables are the most relevant for random forests when making decisions.

With so many variables been fed into the causal forest, it is fairly unlikely that all independent variables are equally relevant for estimating causal effects. The feature importance metric assigns to each independent variable a numeric value that represents how relevant it

⁶There have been many efforts lately to make understandable the inner workings of machine learning models. For a good (and free) overview of the techniques on how to interpret these models, see Molnar 2019.

is for the random forest when making its predictions (Hastie, R. Tibshirani, and Friedman 2009). Feature importance is calculated according to the following description:

- For each node of every tree in the causal forest, determine the variable that splits it.
- For each independent variable, sum up the resulting improvement in the loss function (here: the increase in the difference between treatment and control related to emergency department usage) whenever this variable is used as the splitting criteria.
- Normalize the results in such a way that every feature importance lies between 0 and 1 and their sum equals to 1.

Table 3 shows the feature importance for the ten most relevant variables. The most important variable is “Number of primary care treatable Emergency Department visits” with 16.66%. This type of visits encompasses the conditions that require immediate care but could have been treated in a primary care setting. The second and third most important variables are “Sum of total charges” and “Sum of total Emergency Department charges”, which are the amount expended with hospitals in general and with emergency department only, respectively. Except for “Age in 2009”, all the eight most important variables are related to previous visits to emergency department. Feature importance, however, does not tell us how each feature affects the number of emergency department visits.

	Feature	Importance
1	num_epct_pre_ed	16.66%
2	charg_tot_pre_ed	14.87%
3	ed_charg_tot_pre_ed	14.51%
4	num_edcnp_pre_ed	10.89%
5	num_ne_pre_ed	10.81%
6	num_edcnpa_pre_ed	7.08%
7	age_2009	4.31%
8	num_unclas_pre_ed	2.58%
9	any_hosp_pre_ed_Yes	2.46%
10	any_chron_pre_ed_Yes	2.44%

Table 3: Feature importance of most important variables

6 Final Remarks

After a brief tour on topics such as causality, heterogeneity in treatment effects and decision trees, we managed to apply a causal forest to real-world data. Causal forest is a promising method to uncover heterogeneity in causal effects, specially for large datasets. It applies battle-tested principles developed in machine learning to causal analysis and it has also convenient statistical properties that allow us to assess the certainty of its estimates. For the Oregon Health Insurance Experiment, causal forests seemed to have identified large differences across its participants with respect to the response to Medicaid coverage. However, the uncertainty of the estimates are very high, which does not allow us to neither confirm the existence nor exclude the possibility of heterogeneous treatment effects.

References

- Athey, Susan, Julie Tibshirani, and Stefan Wager (Apr. 2019). “Generalized Random Forests”. In: *The Annals of Statistics* 47.2. Originally circulated as arXiv:1610.01271, pp. 1148–1178. DOI: 10.1214/18-AOS1709. URL: <https://doi.org/10.1214/18-AOS1709>.
- Breiman, Leo (Dec. 1996). “Heuristics of Instability and Stabilization in Model Selection”. In: *The Annals of Statistics* 24.6, pp. 2350–2383. DOI: 10.1214/aos/1032181158. URL: <https://doi.org/10.1214/aos/1032181158>.
- (Oct. 2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324. URL: <https://doi.org/10.1023/A:1010933404324>.
- Finkelstein, Amy et al. (Aug. 2012). “The Oregon Health Insurance Experiment: Evidence from the First Year”. In: *The Quarterly Journal of Economics* 127.3, pp. 1057–1106. DOI: 10.1093/qje/qjs020. URL: <https://doi.org/10.1093/qje/qjs020>.
- Grus, Joel (Apr. 2015). *Data Science from Scratch: First Principles with Python*. 1st ed. O’Reilly Media, p. 330. ISBN: 9781491901427. URL: <https://www.oreilly.com/library/view/data-science-from/9781491901410/>.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York: Springer. ISBN: 9780387848570. DOI: 10.1007/978-0-387-84858-7. URL: <https://doi.org/10.1007/978-0-387-84858-7>.
- Imbens, Guido W. and Joshua D. Angrist (Mar. 1994). “Identification and Estimation of Local Average Treatment Effects”. In: *Econometrica* 62.2, pp. 467–475. DOI: 10.2307/2951620. URL: <https://doi.org/10.2307/2951620>.
- Molnar, Christoph (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com. ISBN: 9780244768522. URL: <https://christophm.github.io/interpretable-ml-book/>.
- Morgan, Stephen L. and Christopher Winship (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Cambridge University Press. ISBN: 9781107694163. DOI: 10.1017/CB09781107587991. URL: <https://doi.org/10.1017/CB09781107587991>.
- Taubman, Sarah L. et al. (Jan. 2014). “Medicaid Increases Emergency-Department Use: Evidence from Oregon’s Health Insurance Experiment”. In: *Science* 343.6168, pp. 263–268. DOI: 10.1126/science.1246183. URL: <https://doi.org/10.1126/science.1246183>.

Wager, Stefan and Susan Athey (July 2018). “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242. DOI: 10.1080/01621459.2017.1319839. URL: <https://doi.org/10.1080/01621459.2017.1319839>.

A Tables

	Code	Meaning
1	age_2009	Age in 2009
2	any_abdo_pre_ed	Any ED visit for abdominal pain, pre-randomization
3	any_acsc_pre_ed	Any ambulatory case sensitive ED visit, pre-randomization
4	any_back_pre_ed	Any ED visit for back pain, pre-randomization
5	any_chron_pre_ed	Any ED visit for chronic condition, pre-randomization
6	any_depres_pre_ed	Any ED visit for mood disorders, pre-randomization
7	any_head_pre_ed	Any ED visit for headache, pre-randomization
8	any_heart_pre_ed	Any ED visit for chest pain, pre-randomization
9	any_hiun_pre_ed	Any ED visit to a high uninsured volume hospital, pre-randomization
10	any_hosp_pre_ed	Any ED visit resulting in a hospitalization, pre-randomization
11	any_inj_pre_ed	Any ED visit for injury, pre-randomization
12	any_loun_pre_ed	Any ED visit to a low uninsured volume hospital, pre-randomization
13	any_off_pre_ed	Any weekend or nighttime ED visits, prerandomization
14	any_on_pre_ed	Any weekday daytime ED visit, pre-randomization
15	any_out_pre_ed	Any Outpatient ED visit, pre-randomization
16	any_psych_pre_ed	Any ED visit for psych conditions/substance abuse, pre-randomization
17	any_skin_pre_ed	Any ED visit for skin conditions, pre-randomization
18	any_visit_pre_ed	Any ED visit, pre-randomization
19	charg_tot_pre_ed	Sum of total charges, pre-randomization
20	ed_charg_tot_pre_ed	Sum of total ED charges, pre-randomization
21	english_list	Individual requested english-language materials: lottery list data
22	female_list	Female: lottery list data
23	first_day_list	Signed up for lottery list on first day: lottery list data
24	have_phone_list	Gave a phone number on lottery sign up: lottery list data
25	last_day_list	Signed up for lottery list on last day: lottery list data
26	num_edcnp_pre_ed	Number of emergent, non-preventable ED visits, pre-randomization
27	num_edcnpa_pre_ed	Number of emergent, preventable ED visits, pre-randomization
28	num_epct_pre_ed	Number of primary care treatable ED visits, pre-randomization
29	num_ne_pre_ed	Number of non-emergent ED visits, pre-randomization(see codebook)
30	num_unclas_pre_ed	Number of of unclassified ED visits, pre-randomization (see codebook)
31	num_visit_cens_ed	Number of ED visits in the study period (Censored)
32	numhh_list	Number of people in household on lottery list
33	pobox_list	Gave a PO Box as an address: lottery list data
34	ret_mode_0m	Method of survey return – 0m
35	sample_0m	Initial mail survey sample (individual was mailed an initial survey)
36	self_list	Individual signed him or herself up for the lottery list
37	surv_lang_0m	Survey language – 0m
38	week_list	Week of lottery sign up: lottery list data

Table 4: Variables used to train the causal forest