Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis

Econometrics and Operations Research

Heterogeneous Treatment Effects of Educational Interventions by using Random Forests

Author:
Fan Fan Jin
Student number:
449599

Supervisor:
Dr. A. A. Naghi
Second assessor:
Dr. A. Pick

July 7, 2019

Abstract

The aim of this paper is to find heterogeneous treatment effects in different applications regarding educational interventions that have not been located before. We shall use a machine learning technique called the Random Forests algorithm implemented by Athey et al. (2019) to reach our goal. With several hypothesis tests and a variable importance measure of the package **grf** in **R**, we can test for heterogeneity and find the most important variables that contribute to this heterogeneity. When implementing this algorithm on the applications of our thesis, we first of all found out that the ability of a student has a high impact on the treatment of camera monitoring on the overall Baccalaureate score in Romania. Also, we found that past math grades and the reading ability have a big importance on the heterogeneity in the application of the grading process in Brazil. Moreover, the number of members and kids in households contribute to the heterogeneity in the normalized math scores of students attending a BRIGHT school in Burkina Faso. All these heterogeneous treatment effects have not been discovered before by past researchers.

Contents

1 Introduction

The transition from high-school to a university is in most countries determined by a final examination. The scores of this examination can be crucial for the further education of a student, because students with higher scores have a higher chance to get into better universities which are better for future careers. This is why students and their parents strive for high grades for these final examinations. Unfortunately, in a lot of countries like Romania, students cheat or commit fraud on these final examinations and will likely get a higher score than they should have. In 2011 the Romanian government used a campaign where they announced that they will monitor the students with CCTV cameras during examination and will also use other punishment treats to target other fraud incidents. This implementation was in reaction to the very high corruption percentage in 2010.

This paper will analyze the effect of the campaign on the fraud incidents of the Romanian Baccalaureate exam, which is the final examination of high-school students in Romania. We shall divide this analysis into two parts. The first part of this paper will consist of a replication of the paper of Borcan et al. (2017), where they already analyzed this effect using a difference-in-difference estimation and looked at some heterogeneity with regards to the students' poverty status. We shall use their published data and code to replicate their results. The second part and the main focus of this paper will contain a different approach, where we shall examine other heterogeneous treatment effects using a machine learning technique called Random Forests studied by Wager and Athey (2018). This technique will combine multiple regression trees, where an individual regression tree is implemented by recursive partitioning. Each split in a tree is made to fully improve the model fit. Several hypothesis tests will be used to determine heterogeneity and a function for variable importance will be applied to find the most important variables for the detection of heterogeneity. Athey et al. (2019) implemented the algorithm and all the other functions in **R** with a package called **grf**. We will use this package in our research and will hopefully get an answer to our main research question: *"Can we find other heterogeneous treatment effects that have not been investigated before by previous researchers using the Random Forests algorithm? And if so, which ones?"*.

We will use the Random Forests algorithm on two other instances to also answer the main research question. Our first application will be the work of Botelho et al. (2015), where they studied the heterogeneous effects of students' race on math grading in Brazil. This application does not include a treatment indicator. Moreover, we will apply the algorithm on the data set of Kazianga et al. (2012), where they found an increase of the test scores of students' attending a BRIGHT school in Burkina Faso, which stands for schools that are more friendly for girls.

Existing knowledge like Borcan et al. (2017) and Botelho et al. (2015) focused their heterogeneous effects solely on one characteristic without using any machine learning techniques and Kazianga et al. (2012) did not even look at any heterogeneity. This is where our research becomes interesting. First of all, machine learning techniques are much more efficient. If you would want to study other heterogeneous treatment effects in the data set used by Borcan et al. (2017) without machine learning, you have to manually check all the 44 variables. Also, there is the chance of the multiple hypothesis testing issue if you do not use machine learning

techniques, which occurs when testing multiple hypotheses simultaneously. Furthermore, the Random Forests algorithm works great for data sets where the number of covariates is large in comparison with the sample size. In these cases OLS will break down, but machine learning techniques still perform well. Moreover, the Random Forests algorithm studies the heterogeneity issue in a much more detailed way than previous researches have done by focusing on only one characteristic.

Machine learning techniques, such as Random Forests, have been in the computer science industry for decades. However, economics could not use these techniques before, because the asymptotic theory was not yet developed in that time. They could predict the outcome of an observation based on their features and characteristics, but they could not obtain significance or other coefficients that were relevant. When time went on, a lot of researchers (Athey et al. (2019), Wager (2014), Scornet (2016)) developed this asymptotic framework for these techniques, such that they are now applicable for economists to use. What we are technically doing is to find out how these machine learning tools (in our case the Random Forests algorithm) perform in economics.

With the extension of the Random Forests algorithm we desire to provide a better comprehension of the heterogeneous treatment effects of several applications. The Romanian government and other policy makers will get more insights for whom the campaign did work to reduce the corruption during the examination. For example, if the algorithm finds out that there is a heterogeneous treatment effect among the sex of a student saying that boys are more prone to cheating than girls then with this result the government can implement more regulations against boys. Moreover, the results for the heterogeneous effects in grading math tests will be relevant for Brazilian teachers and Brazilian policy makers of education, because they will have a detailed answer on the differences between math scores. Furthermore, policy makers of Burkina Faso and founders of the BRIGHT schools get to know the most important characteristics of the students that are more likely to have a higher test score when attending a BRIGHT school.

When implementing the Random Forests algorithm to the research of Borcan et al. (2017), we indeed found other heterogeneous treatment effects than the one they already stated. We found out that variables regarding the ability of a student have a high importance in the detection of heterogeneous treatment effects. Moreover, in our second application of Botelho et al. (2015) we also found other heterogeneous effects that have not been studied before by past researchers. In this application we found out that past math grades and the reading ability of a student have a high importance in detecting heterogeneous effects in the present math grades, while previous researchers only looked into differences in race. At last, in the data set of Kazianga et al. (2012) we again found other heterogeneous treatment effects that are new to the literature. We found out that the number of members and kids in a household are important for the heterogeneous treatment effects of BRIGHT schools on the normalized math scores of students.

Furthermore in this thesis you can find a literature review in section 2. An explanation of the data follows in section 3. Moreover, in section 4 you can find a description of the methods used in our research. The results will be presented in section 5. Finally, we shall end the thesis with a conclusion in section 6.

2 Literature

To replicate the paper of Borcan et al. (2017), we use a difference-in-difference estimation that has been in the literature for decades. Lechner (2011) gives a brief overview of all the literature that has been done about this estimation method for causal inferences and also summarizes the different issues that occurred. For example, a consistent estimator does not exist when using an estimation with a finite number of periods. The DID-estimation is mostly used to estimate effects of certain public interventions (e.g., Conley and Taber (2005)), but it can also be used in social sciences (e.g., Gebel and Voßemer (2014)). In addition, just as in our research, the DID-estimation is mostly used in linear models. However, Ai and Norton (2003) investigated the DID-estimation in non-linear models such as probit and logit models, because in non-linear models you can not calculate the statistical significance by using standard software. Also, panel data can be applied to such estimation strategies, which are called individual-level panel data. Furthermore, there are several semiparametric and nonparameteric approaches (Athey and Imbens (2016)).

For the main part of this paper we shall use the Random Forests algorithm. A lot of researchers (Athey et al. (2019), Wager (2014), Scornet (2016)) developed the asymptotic framework for these techniques, such that they are now applicable for economists to use. Breiman (2001) proposed the original Random Forests algorithm applicable for non-parametric classification. Athey et al. (2019) generalized this algorithm and developed new techniques for statistical tasks such as the estimation of heterogeneous treatment effects via instrumental variables. This algorithm has been applied by several researchers so far (e.g., Knaus et al. (2018), Korepanova et al. (2019), Viviano and Bradic (2019)).

The Random Forests algorithm can also be combined with other machine learning techniques, like Oprescu et al. (2018) studied. They proposed an orthogonal random forest which uses Neyman-orthogonal moments with the generalized random forests algorithm of Athey et al. (2019). With this approach they also estimate heterogeneous treatment effects and showed that this approach controls for an even higher dimensional set of variables than the approach stated without orthogonal moments. By combining these two machine learning techniques they also reduce sensitivity with regards to the parameters to estimate the target parameter.

However, the Random Forests algorithm is not the only way to analyze heterogeneous treatment effects. Athey and Imbens (2016) analyzed these effects by recursive partitioning. They estimated this effect with an honest approach, where one part of the data is used to construct the separation and the other part to estimate the treatment effects for each subgroup. After this they also built a regression tree to optimize the goodness of fit for the treatment effects. We can say that Athey and Imbens (2016) only used one individual regression tree for their analysis, while we will work with a forest of B such trees. Also, Wang et al. (2017) even found out with simulation that his approach using instrumental variable trees outperforms the generalized random forest. But since we do not make use of instrumental variables we stick to the generalized random forest by Athey et al. (2019).

3 Data

This section will explain all the data we have used in our three applications.

3.1 Romanian Baccalaureate Scores

The data we shall use in our first application of our research will be obtained from the paper of Borcan et al. (2017). We will also use this data set to replicate the work of Borcan et al. (2017) using a difference-in-difference estimation. They used two data sets from different sources for their investigation of the effect of CCTV camera monitoring on the Baccalaureate scores and combined the sources after they secured all the data.

First of all, administrative data, such as students' exam score, gender, date of birth, county etc., for the years 2009 till 2012 is provided by the Ministry of Education. This data set covers all students enrolled for the Baccalaureate exam. For 70% of the students we also have the average final scores in middle school. This data will be used as a control for ability.

Secondly, the students' poverty status to examine the heterogeneity issues is obtained from the source of the Money for High School (MHS) public program of financial assistance. High-school students living in a household with a short of money were eligible for this program. A household is stated poor if the gross monthly income per household member did not exceed 150 RON (US $45). This data set consists of dummy variables, where one stands for students that were eligible for the program in a specific year and zero otherwise.

These two data sets from the Ministry of Education and the MHS program combined lead us to 731,505 students and 44 variables, including indicators for every year and every county. Re-examinations were not taken into account in this data set, which means for every student in our data set we only observe their exam scores once.

3.1.1 Descriptive statistics

The descriptive statistics of the data set sorted per year are stated in the table below. The first three variables are the different outcomes of the Baccalaureate exam. The written examination score is the score on a scale of 1 to 10 for the written exam of the Romanian language. The examination pass is a dummy variable where one stands for that the student passed the Baccalaureate exam and zero stands for a fail. And the overall examination score is also a grade on a scale of 1 to 10, which stands for the average score in the overall Baccalaureate exam. We immediately note for these three variables that the mean in the years 2011 and 2012 are much lower than in 2009 and 2010. This can indicate that the campaign to reduce corruption at the end of 2010 has probably worked out. The table also shows a slight increase in the fraction of students with a poor poverty status, while the sample size decreases over the years. The decrease of observations in 2012 may be due to the lack of registrations for the Baccalaureate exam. The features of gender, theoretical track, rural and low ability also have slight changes over time, but in general they are quite continuous.

Table 1: Descriptive statistics of the data set of Borcan et al. (2017) per year.

	2009		2010		2011		2012	
Variables	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Written Examination Score	6.813	1.819	7.017	1.664	6.147	2.102	6.143	2.138
Examination Pass	0.813	0.390	0.692	0.462	0.482	0.500	0.482	0.500
Overall Examination Score	8.057	1.150	6.969	1.647	6.033	1.998	6.049	2.142
Poor	0.166	0.372	0.175	0.380	0.185	0.388	0.201	0.401
Male	0.483	0.500	0.490	0.500	0.480	0.500	0.463	0.499
Theoretical Track	0.447	0.497	0.434	0.496	0.447	0.497	0.469	0.499
Rural	0.057	0.232	0.065	0.246	0.067	0.250	0.059	0.236
Low Ability	0.509	0.500	0.514	0.500	0.500	0.500	0.468	0.499
Total Observations	196,687		195,755		182,939		156,124	

Note: SD stands for the standard deviation. Source by: Borcan et al. (2017) (table 1).

3.2 Math grading in Brazil

For the second application on the Random Forests algorithm, we shall use the data set of Botelho et al. (2015) to analyze heterogeneous effects in math grading for the country Brazil. They focused their research on the eighth graders in the city of Sao Paulo in the year 2010.

From the Secretary's data bank they attained four variables obtaining several records and proficiency's of the students. Also, the administrative data set of the teachers' evaluations on all the students will be used to examine the grading discrimination. This data set contains a lot of information about grades and attendance records.

Moreover, students, parents and teachers were asked several questions regarding to demographics, race, study habits and economic status to obtain more variables in the data set. Using standardized scores of Sao Paulo's Performance Evaluation System (SARESP), they merged all these different data sets together and obtained a working sample of 277,444 students across 3,511 different schools and 89 variables. This data set is restricted by non-homogeneous classrooms with regards to race and a classroom needs to consist of at least 15 students.

3.3 BRIGHT schools in Burkina Faso

The last application we shall apply on the Random Forests algorithm will be the data set of the work of Kazianga et al. (2012). Using this data set we will analyze the heterogeneous treatment effects of BRIGHT school on the test scores of the students.

BRIGHT schools stand for friendly schools for girls that have more amenities than the average school in the rural areas of Burkina Faso, such as separate toilets per gender, a special training program for mothers and teachers got offered a special training for gender sensitivity. All the data used for this application is obtained from a survey conducted in 287 villages. All households with children attending primary school in one of the 287 villages were conducted. The survey consisted of several socio-demographic questions, test scores of the childrens' math and French ability and attendance records of the children. In total this data set consists of 115 variables and 23,282 observations.

4 Methodology

In this section we will explain the methods we will use in our research.

4.1 Difference-in-difference estimation

First of all, for the replication of the paper of Borcan et al. (2017) we will use a difference-in-difference (DID) estimation. Ever since the research by Ashenfelter and Card (1984), the use of this estimation has become very popular among researchers. The DID strategy is perfect applicable for calculating treatment effects without randomization. There are two groups and two time periods in the general form of the DID estimation, namely the treatment group and the control group. The observations in the treatment group will be subjected to some intervention, whereas the observations in the control group will not be subjected to the intervention. This approach will compare the differences in outcomes between the two groups, which will be the DID estimate. The estimation of coefficients are computed by OLS.

The DID estimation can also be applied to multiple time periods, which will be referred to as a multilevel model. The multilevel model will take into account the time trend by eliminating the influences of this time trend using the control group. Both groups need to have the same time trend in absence of treatment. This is also called the common trend assumption.

The specification of the DID estimation of Borcan et al. (2017) is stated as follows:

$$y_{ict} = \alpha + \beta T_{ct} + \gamma' X_{ict} + \phi_t + \theta_c + \theta_c \cdot t + \varepsilon_{ict}, \tag{1}$$

where i stands for a student in county c in year t. The dependent variable y_{ict} is one of the three types of scores of the Romanian Baccalaureate exam. It can be the score of the written Romanian language test, a dummy variable that indicates one if the student passed the Baccalaureate or the full score of the Baccalaureate. The variable T_{ct} is also a dummy variable that specifies if a student is monitored by a CCTV camera in county c in year t. X_{ict} includes indicators stating if a student i from county c in year t comes from a poor status, the sex of a student, their school track and if the student is from a rural area. The ϕ_t expresses year indicators for each year t and θ_c includes the 41 county indicators for each county c.

Since we use multiple time periods in the specification stated in equation 1, we are working with a multilevel difference-in-difference model. The treatment group in our case is monitored by a CCTV camera and the control group is not monitored by a CCTV camera. This specification also applies the common trend assumption with the variable $\theta_c \cdot t$. The difference-in-difference estimate is in our specification the $\hat{\beta}$ and will express the impact of the CCTV monitoring on the exam scores taking into account the variation within the different counties over time. We shall also look at heterogeneity between poor and non poor students.

We will use the code published by Borcan et al. (2017) to replicate some of their results that are interesting for our research. For the full code used in our research we refer to the Appendix at the end of this thesis.

4.2 Random Forests

To be able to find more heterogeneous effects in the data set, we will implement the Random Forests algorithm by Athey et al. (2019). We shall take advantage of the developed software package **grf** in **R**. For a full code of the algorithm and its applications we refer to the Appendix at the end of this thesis. In this section we shall explain the meaning and purpose of a causal tree, a causal forest and we shall further explain the testing of heterogeneity and the extra applications on this algorithm.

4.2.1 Causal tree

We would like to examine the treatment effect given some feature x for every observation i, we can introduce the treatment effect as stated in Athey et al. (2019):

$$\tau_i(x) = \mathbb{E}[Y_i^{(1)} - Y_i^{(0)} \mid X_i = x], \tag{2}$$

where $Y_i^{(1)}$ and $Y_i^{(0)}$ are the outcomes with and without treatment for the observation i, respectively. We note the treatment indicator as $W_i \in \{0, 1\}$. However, if an individual i is treated by a treatment indicator it can not be not treated at the same time, thus we do not observe this treatment effect. In other words, we can only observe $Y_i^{(1)}$ or $Y_i^{(0)}$, but we can not observe both of them simultaneously. Therefore, we assume unconfoundedness, which can be mathematically denoted as:

$$\{Y_i^{(0)}, Y_i^{(1)}\} \perp W_i \mid X_i. \tag{3}$$

This notation is taken from the work of Athey et al. (2019). Unconfoundedness means that W_i, the treatment effect, is independent of the outcome Y_i conditional on the feature X_i. This implies that it is possible to assume that the observations nearby the x-space are coming from a randomized trial with $x \in X$. Therefore, local methods are consistent for $\tau_i(x)$. Observations that are nearby in the x-space are now in the context of decision trees defined as observations that are in the exact same leaf as x. This means that in each leaf we can say that the observations come from a randomized trial.

To build a causal tree we start by partitioning the feature space into a set of leaves L. Each split is chosen such that it maximizes the improvement to fit the model, which means we would like to maximize the heterogeneity. The first split will be made of the most important feature and the second split of the second most important feature etc., such that you can see in a much more detailed way where the heterogeneity is coming from. With this in mind we can estimate the treatment effect in the context of trees for any $x \in L$ as follows:

$$\hat{\tau}_i(x) = \frac{1}{|\{i : W_i = 1, X_i \in L(x)\}|} \sum_{\{i:W_i=1, X_i \in L(x)\}}^{Y_i} - \frac{1}{|\{i : W_i = 0, X_i \in L(x)\}|} \sum_{\{i:W_i=0, X_i \in L(x)\}}^{Y_i}. \tag{4}$$

This equation has also been taken from the paper of Athey et al. (2019).

4.2.2 Causal forest

Until now we described single causal trees, but in our research we would like to use a random forest. Athey et al. (2019) describes several different forests for different applications in their work. However, in the context of the work of Borcan et al. (2017) the most applicable approach is the causal forest, since we are studying the causal effects of a policy intervention. A causal forest actually generates a group of B single causal trees with each $b \in B$ a corresponding estimate of $\hat{\tau}_b(x)$. The causal forest aggregates all these predictions and takes the average of them: $\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$, which calculates the average treatment effect. In the figure below is shown how this works graphically for 600 trees.

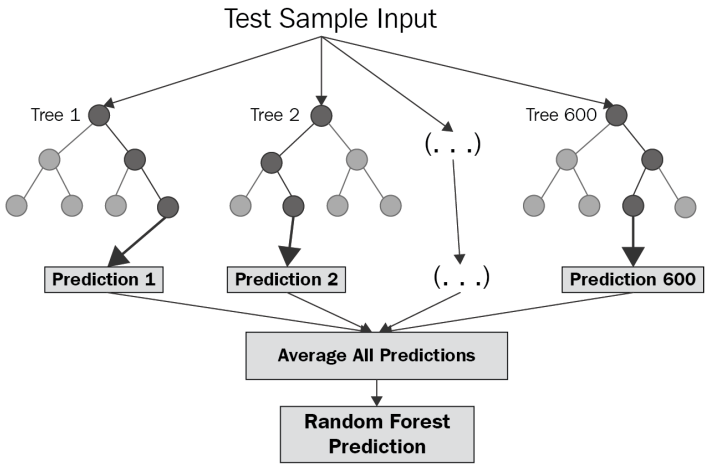

Figure 1: Random Forest algorithm.
Source by: https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789132212/2/ch02lvl1sec21/decision-tree-based-ensemble-methods

Each tree that is used to build the causal forest needs to be implemented by a different training sample of the data set. The training sample is necessary to fit the model and parameters and is a subset of the data. Also, each tree in the causal forest needs to be honest. A causal tree is honest if the outcome Y_i is the only parameter that is used to estimate the treatment effect τ within a leaf for each individual i in the training sample using equation 4. Breiman (2001) showed that it is better to generate different trees and averaging their predictions, because this helps to reduce the variance of the model. It is also really difficult to find a single highly-optimized causal tree.

However, Athey et al. (2019) showed another approach to estimate the parameter of interest $\hat{\tau}(x)$. They defined similarity weights $\alpha_i(x)$ that evaluate the relevance of the i-th training example, which is an observation in the training sample, to fit the parameter of interest at x or in other words captures the frequency of which training example i falls into the exact same leaf as x. Assuming a forest is made out of B trees. The set of training examples that fall in the exact same leaf as x for each $b \in B$ will be denoted as $L_b(x)$. Athey et al. (2019) denoted the weights $\alpha_i(x)$ as:

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \alpha_{b_i}(x) \quad \text{with} \quad \alpha_{b_i}(x) = \frac{\mathbf{1}(\{X_i \in L_b(x))\}}{|L_b(x)|}. \tag{5}$$

The weights sum up to 1. After obtaining the weights we can calculate the average treatment effect taking into account the importance of the different training examples used to generate the trees in the forest. We can say that we seek for trees in our forest that cause weights $\alpha_i(x)$ that lead to better estimates of $\hat{\tau}(x)$, where $\hat{\tau}(x)$ is now calculated as: $N^{-1}\big(\sum_{i=1}^{N} \hat{\tau}_{b_i}(x)\alpha_i(x)\big)$.

The predictions of the treatment effect for every observation i of a causal forest can also be computed with a function in the package **grf**. There are two ways to predict these treatment effects. The first way is to use out-of-bag predictions. This approach provides predictions using trees that did not use the i-th observation in their training sample. The second way is to use a test sample to predict these estimates. A test sample can not be confused with a training sample, because the main purpose of a training sample is to perfectly fit the model, while the test sample is used to validate and predict the model. Breiman (1996) states that the out-of-bag estimation is just as accurate as applying a test sample of equal size as the training sample. Therefore, we shall use out-of-bag predictions for our research.

We can define the Random Forests algorithm in a few steps. Athey et al. (2019) have stated this algorithm in their paper as shown below.

Algorithm 1 Generalized random forest with honesty and subsampling

All parameters are pre-specified, such as the amount of trees B and the sub-sampling rate s.

procedure GENERALIZED RANDOM FOREST(set of training examples $\mathcal{S}$, point x)
  weight vector $\alpha \leftarrow$ ZEROS($|\,\mathcal{S}\,|$)
  for $b = 1$ till B **do**
    set of examples $\mathcal{I} \leftarrow$ SUBSAMPLE($\mathcal{S}, s$)
    set of examples $\mathcal{J}_1, \mathcal{J}_2 \leftarrow$ SPLITSAMPLE($\mathcal{I}$)
    tree $\mathcal{T} \leftarrow$ GRADIENTTREE($\mathcal{J}_1, \mathcal{X}$)
    $\mathcal{N} \leftarrow$ NEIGHBORS(x, $\mathcal{T}, \mathcal{J}_2$)
    for all example $e \in \mathcal{N}$ **do**
      $\alpha[e]$ += $1/|\,\mathcal{N}\,|$
    end for
  end for
  output $\hat{\theta}(x)$ with weights α/B

The default parameters of this algorithm in **grf** are: B is 2000 and s is 50%. For an explanation on the purpose of the different functions within this algorithm, we refer to the paper of Athey et al. (2019). Source by: Athey et al. (2019) (algorithm 1).

For the asymptotic theory and analysis we refer to the work of Athey et al. (2019), since they provide a very elaborate description.

Using a Random Forest brings a few big advantages to the research, since it reduces the change of over-fitting the data and it can be used for classification and regression. However, the computation and prediction process do require much more time than a single decision tree. Random Forests are also known to handle missing values quite well in a data set. Unfortunately, the **grf** package used for our work is not resistant for these missing values.

4.2.3 Testing and finding heterogeneity

The causal forest can predict the treatment effects of every observation i and the average treatment effect of the whole working sample. However, these estimates do not tell if the causal forest has successfully detected heterogeneity in the treatment estimates. In most analyses the treatment effects can variate among the observations. For this purpose we will use two different approaches to test for heterogeneous treatment effects.

First of all, a simple approach is to divide the out-of-bag predictions into two groups and test if the average treatment effect of these two groups differ from each other. When predicting the treatment effects of all the observations, the **grf** package will give a vector as output with for every observation i a prediction of τ_i calculated as in equation 4. We shall divide these two groups as explained in Athey and Wager (2019). The first group will consist of observations with an out-of-bag prediction larger than the median of all the predictions and the second group will consist of the observations with an out-of-bag prediction smaller than this median. Lets denote the average treatment effect of the two groups as $\hat{\tau}_g(x)$ with $g = 1, 2$ respectively for the first and the second group. Using the notation of the average predicted treatment effect of the two groups, we can introduce the difference of the average treatment effect as follows:

$$\hat{\tau}_{diff}(x) = \hat{\tau}_1(x) - \hat{\tau}_2(x). \tag{6}$$

The null hypothesis for this test is that the average treatment effect of both groups are equal to each other, $H_0 : \hat{\tau}_1(x) = \hat{\tau}_2(x)$ or that the difference between these two estimates is equal to 0, $H_0 : \hat{\tau}_{diff} = 0$. The alternative hypothesis will then obviously be, $H_1 : \hat{\tau}_1(x) \neq \hat{\tau}_2(x)$ or $H_1 : \hat{\tau}_{diff} \neq 0$. We shall use a 5% significance level to evaluate the result of this test and if we can reject the null hypothesis we can conclude that there are indeed heterogeneous treatment effects in our research.‘

The second approach we shall use is a test motivated by Chernozhukov et al. (2017). This approach aims to fit the average treatment effect as a linear function of the out-of-bag predictions and is also provided in the **grf** package referred as the calibration test. The function in this package will apply a heteroskedasticity-consistent test of calibration and will give the estimates, standard error, t-statistic and p-value of two predictions. These two predictions are the *mean.forest.prediction* and the *differential.forest.prediction*. We can use the *differential.forest.prediction* to test for the existence of heterogeneity. If the corresponding p-value of this coefficient is smaller than the significance level of 5%, then we can reject the null hypothesis of no presence of heterogeneity.

If we can reject the null hypothesis in both approaches, then the next step is to determine which features are the most important for the detection of heterogeneity. We will again use a function in the **grf** package that gives us the importance of each feature $x \in X$ used in the causal forest. In most cases the most important feature will be the feature that has been split on in one of the first stages of the trees. For each split of each tree, the causal forest evaluates the impact of heterogeneity that each variable causes. It will split each tree on the variable that maximizes the heterogeneity. Hence, the variable importance function in the **grf** package measures how often the causal forest selects each feature to grow on its trees. The function will

give a percentage for each feature $x \in X$. When ordering this vector from high to low, we can find the top 5 most important variables for detecting heterogeneous treatment effects.

When obtaining the most important variables for the heterogeneity, we can calculate the heterogeneous treatment effect of these variables. This will be done in an almost similar way as the first approach of testing for heterogeneity. We shall again divide the sample into two groups, where one group will consist of all observations with a value higher than the median of all the values and the other group will consist all the other observations with a value lower than the median. This will be done for each of the most important variables in our analysis for detecting heterogeneous treatment effects. We will not use the predicted treatment effects, but the data we have for each variable. For each group (above and below median) we will calculate the average treatment effect. Lets denote these estimates as $\hat{\tau}_{high}$ and $\hat{\tau}_{low}$ respectively. To double check if this variable is indeed of importance for the heterogeneity, we can use a t-test to test the difference between $\hat{\tau}_{high}$ and $\hat{\tau}_{low}$. This difference between the two estimates is referred to as the heterogeneous treatment effect.

4.2.4 Applications

We can translate all the theory explained above into the context of our three applications.

First of all, as stated before the **grf** package is not applicable for missing values in the data set obtained from Borcan et al. (2017). Thus, the first step is to omit all the observations with a missing value in one of the features in our first data set. After deleting all these observations we go from 731,505 students that took the Baccalaureate exam in our working sample to 100,604 students. However, due to the lack of memory capacity on laptops and computers we filter these 100,604 students with regards to their student number to get a smaller working sample. In our working sample we choose students with a student number smaller or equal than 800,000 of those 100,604 students. This leads us to a final working sample of 47,909 students with a student number smaller or equal than 800,000 and without any missing values.

When observing all the variables in our data set the variable $Y_i^{(W_i)}$ is one of the outcomes of the Baccalaureate exam: the score of the written examination, the dummy stating if a student passed the Baccalaureate or the overall examination score. The treatment effect W_i will be the dummy variable stating if the corresponding student i got monitored by a CCTV camera or not. For 2011 this was the case for 25 counties, whereas for 2012 all students of all the counties in Romania were monitored by the CCTV cameras. The X_i is one of the 44 variables in our data set indicating some feature, such as poverty status, sex or school track. Finally, the $\hat{\theta}(x)$ in our case is the average treatment effect, since that is our parameter of interest.

Another approach we will add to our causal forest are clusters (Abadie et al. (2017)). Since each student is from a different county, we can cluster our causal forest based on the different counties. The treatment effect (CCTV camera monitoring) is different among the counties, because in 2011 only 25 counties got subjected to the treatment while in 2012 all counties were subjected to the treatment. This opens up the question that there might be heterogeneity across counties. Also, we would like results that generalize beyond the counties. This means that we want models that can predict the effects of a new student with its features from a new county as accurate as possible.

Working with clusters in the causal forest do require some adaptions from the Random Forests algorithm as stated in algorithm 1. In line 4 where the function SUBSAMPLE generally draws a subsample of students, will now draw a subsample of clusters. Also, when making out-of-bag predictions using clusters will consider a student i to be out-of-bag when the county where the student belongs to was not drawn to be in the training sample.

Our second application will not be applied to a causal forest, but on a regression forest. Botelho et al. (2015) did not use a treatment variable, but used a regression to estimate the dependent variable, the z-scores of the Brazilian math grades in 2010. The regression forest can also be implemented by the **grf** package of Athey et al. (2019). However, it does have its differences in comparison with a causal forest. Each causal tree will have its weights as denoted in equation 5. Nevertheless, a regression tree will have its weights equal to the approach stated in Breiman (2001), which will be the average of all the predictions made by each $b \in B$ stated as $\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$. Also, the interpretation of the predictions of the regression forest are different than the predictions of the causal forest. The causal forest will predict a treatment effect for each observation i, but since regression trees do not consist of a treatment variable the regression forest can not predict this. The regression forest will predict estimates for the dependent variable, mathematically denoted as: $\mathrm{E}[Y|x = X]$. This implies we will use the hypothesis test using the difference between the estimated dependent variable of the two groups above and below the median for identification of heterogeneous effects.

Variable importance and the calibration test can all be applied to the regression forest. Also, the regression forest implemented in the **grf** package is just like the causal forest not resistant for missing values. However, the data set published by Botelho et al. (2015) does not contain any missing values. Due to lack of capacity on laptops and computers, we shall use the first 50,000 observations out of the 277,444 observations to implement the regression forest on. The covariates in the X matrix will consist of 88 variables, where we do not use the dependent variable, z-scores of the math grades in 2010, in our covariates matrix.

For our third application we will use the causal forest just as our first application, since there is a treatment effect included in the data set of Kazianga et al. (2012). The treatment group in the third application are all students attending a BRIGHT school, because we would like to find heterogeneous treatment effects of students attending BRIGHT schools on the test scores of the math test. The outcome or the dependent variable will be the normalized math test score of each student and the treatment variable will be a vector of 1 if a student goes to a BRIGHT school and 0 otherwise. Our data set consists of 23,282 observations and 115 variables. However, some variables have a lot of missing values and some do not have a numeric value, such as the department of a student or the name of a student. We delete all these variables, because including the variables with a lot of missing values will give us a really small working sample. Results using this small working sample will not be reliable. Also, the **grf** is not resistant for missing values and non-numeric values. Therefore, we are obliged to only use variables with numeric data and a decent sized working sample without any missing values. After all these implementations we finalize our working sample of 17,664 observations and 90 variables. All the steps in the Random Forests algorithm will be the same as the application of Borcan et al. (2017). The only difference is that using clusters in not applicable for this application.

5 Results

In this section you can find all the results obtained from the difference-in-difference estimation and the Random Forests algorithm.

5.1 Difference-in-difference estimation

The main results without looking at any heterogeneity of the difference-in-difference estimation in equation 1 are shown in table 2. The first column of each outcome is only regressed on the year indicators with 2010 as the base year. Moreover, the second column of each outcome included the CCTV camera monitoring. Furthermore, the third column adds the control variables: poverty status, gender, theoretical track and the different rural indicators. And the last column also includes county specific linear time trends. This regression regressed the three different Baccalaureate outcomes on the treatment intervention of camera monitoring and the year dummies. The control variables include the poverty status, the gender, the theoretical track and the different rural indicators.

We note in all the first columns of each outcome that the estimate for the years 2011 and 2012 have a sharp decrease with regards to the base year of 2010. When including the CCTV monitoring treatment effect in the regression, we also find a decrease of all the outcomes due to the CCTV camera's monitored in 2011 and 2012. There is not a lot of change between the second and the third column of all the three outcomes implying that adding the control variables stated above does not change the interpretation of the estimates. However, when we add the county specific linear time trends the magnitude of the effect of the camera treatment is slightly larger than without the county specific linear time trends. For a more elaborate table of table 2 we refer to the Appendix at the end of this paper.

Table 2: Main results of the DID-estimation per outcome.

	Written Examination Score				Examination Pass				Overall Examination Score			
Camera		-0.246	-0.251	-0.353		-0.076	-0.076	-0.095		-0.430	-0.439	-0.512
		(0.108)	(0.106)	(0.106)		(0.03)	(0.029)	(0.025)		(0.144)	(0.142)	(0.137)
2012	-0.874	-0.628	-0.716	-0.463	-0.211	-0.135	-0.148	-0.082	-0.923	-0.492	-0.579	-0.323
	(0.065)	(0.087)	(0.078)	(0.087)	(0.024)	(0.025)	(0.023)	(0.017)	(0.092)	(0.115)	(0.106)	(0.094)
2011	-0.875	-0.713	-0.743	-0.597	-0.211	-0.161	-0.166	-0.129	-0.943	-0.660	-0.690	-0.547
	(0.058)	(0.07)	(0.071)	(0.081)	(0.022)	(0.019)	(0.018)	(0.016)	(0.088)	(0.091)	(0.09)	(0.088)
2009	-0.205	-0.205	-0.237	-0.311	0.121	0.121	0.115	0.093	1.087	1.086	1.055	0.967
	(0.054)	(0.054)	(0.057)	(0.033)	(0.011)	(0.011)	(0.011)	(0.012)	(0.042)	(0.042)	(0.04)	(0.04)

Note: the corresponding standard error is shown in brackets. Source by: Borcan et al. (2017) (table 2).

To analyze heterogeneous effects in poverty status we also use equation 1 to get the regression estimates, but in this case the coefficients will differ among the poor and the non-poor students. These results are shown in table 3. This regression includes county fixed effects, year fixed effects, county yearly trends and the control variables as explained above. The first column of each outcome presents the coefficients for the full working sample. The second column shows the same estimates as in the first column, but only on a subset of the working sample. This subset contains all observations with a value for the variable ability. This regression is done to

better comprehend the different outcomes between the poor and the non-poor students. The last column of each outcome also includes ability interactions in the regression analysis.

We note that the estimates of the third columns do not differ that much with regards to the estimates of the second columns, indicating that students from the same ability can have different outcomes on the exam. This together with the estimates of the first column of each outcome tells us that the campaign caused more inequality for the poor students with the same ability. Also, when observing the first and the second row with each other, we note for almost all cases a slight difference. This implies that there are indeed heterogeneous treatment effects among the poverty status of students. For a more elaborate table of table 3 we refer to the Appendix at the end of this paper.

Table 3: Heterogeneous effects by using the poverty status of students.

	Written Examination Score			Examination Pass			Overall Examination Score		
Camera	-0.302	-0.255	-0.131	-0.081	-0.088	-0.077	-0.433	-0.434	-0.308
	(0.110)	(0.082)	(0.056)	(0.026)	(0.023)	(0.017)	(0.141)	(0.113)	(0.076)
Poor × Camera	-0.213	-0.257	-0.214	-0.062	-0.056	-0.051	-0.346	-0.350	-0.306
	(0.063)	(0.053)	(0.052)	(0.015)	(0.012)	(0.013)	(0.078)	(0.061)	(0.063)
Observations	712,298	547,447	547,447	731,505	553,903	553,903	706,895	545,121	545,121

Note: The corresponding standard error is shown in brackets. Source by: Borcan et al. (2017) (table 4).

5.2 Random Forests

We shall present our results in this section using the **grf** package in **R** on our three applications of Borcan et al. (2017), Botelho et al. (2015) and Kazianga et al. (2012), respectively.

5.2.1 Romanian Baccalaureate scores

We implemented the Random Forests algorithm of the Romanian Baccalaureate scores on three different outcomes. The average treatment effects with its corresponding standard error is shown in table 4 for each of the outcomes. We can see that the average treatment effect of -0.289 indicating that the students who got monitored by the cameras have on average an overall examination score of 0.289 lower than the students who were not monitored.

Table 4: Average treatment effect estimates per outcome.

	Overall Examination Score	Examination Pass	Written Examination Score
Average treatment effect	-0.289 (0.080)	0.016 (0.014)	0.018 (0.029)

Note: The standard error is given in brackets under the corresponding estimate.

In table 5 are the estimates shown of the calibration test studied by Chernozhukov et al. (2017) for the coefficient *differential.forest.prediction* with its corresponding p-value and standard error and the t-test for $\hat{\tau}_{diff}$ as explained in section 4.2.3. We can observe that only the p-value for the overall examination score is smaller than our significance level 0.05 using the calibration test and for the t-test all outcomes have a p-value smaller than 0.05. Combining

these two results means that we can only reject the null hypothesis of no heterogeneity for the overall examination score. For the two other outcomes we do not reject the null hypothesis of no heterogeneity.

Table 5: The results of the hypothesis tests for heterogeneity per outcome.

	Overall Examination Score			Examination Pass			Written Examination Score		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value
Calibration test	1.633	0.464	0.000***	-1.447	1.439	0.843	-3.560	1.808	0.976
	t-value		p-value	t-value		p-value	t-value		p-value
t-test	146.490		0.000***	193.470		0.000***	133.180		0.000***

Note: The significance code *** stands for a strong significance.

Table 6 shows which variables are the most important for the detection of heterogeneity in the data set for the outcome of the overall examination score. We now only focus on this outcome, since we could not reject the null hypothesis of no heterogeneity for the other outcomes. We note that the percentile ranking of the written examination score is one of the most important features. Remarkable is that the poverty status of the students, studied by Borcan et al. (2017), is not in the top five of the most important variables for heterogeneous treatment effects. Out of these five variables, we shall find the heterogeneous treatment effect for the ranking of the written examination, the mid school GPA and the ability, because the other two variables are dependent variables in our analysis.

Table 6: The importance of the five most important variables for heterogeneity.

Variable	Overall Score	**Ranking Written Exam**	Written Exam Score	**Mid school GPA**	**Ability**
Importance	23.0%	21.6%	21.3%	4.5%	4.4%

Note: These results are made for the dependent variable of the overall examination score on the Romanian Baccalaureate.

The heterogeneous treatment effects of the most important variables are shown in table 7. We also analyze the heterogeneous treatment effect for the poverty status to see if the result is similar to the result of the difference-in-difference estimation. We find a similar result for the ranking of the written examination, the mid school GPA and the ability. First of all, we can reject the null hypothesis of no heterogeneity for all variables. Also, when observing the $\hat{\tau}_{low}$ with the $\hat{\tau}_{high}$ we find a much larger magnitude of decrease in the dependent variable for the group with values below the median. This can indicate that the higher a student is ranked the percentile ranking of the written examination, the higher the student will score in the overall examination score of the Baccalaureate. For the analysis of the poverty status' we find a similar result as in table 3. In table 7 we find an average predicted treatment effect of -0.260 for the poor students, while in table 3 we found an estimate of -0.213 for the camera intervention with an interaction with the poverty status.

Table 7: Heterogeneous treatment effects of the most important variables.

Variables	$\hat{\tau}_{high}$	$\hat{\tau}_{low}$	t-value	p-value
Ranking written exam	-0.053	-0.333	117.610	0.000
	(0.0004)	(0.001)		
Mid school GPA	-0.085	-0.309	88.987	0.000
	(0.0004)	(0.001)		
Ability	-0.083	-0.311	90.386	0.000
	(0.0004)	(0.001)		
Poor	-0.260	-0.176	-24.455	0.000
	(0.001)	(0.001)		

Note: The corresponding standard errors are shown in brackets.Please refer to section 4.2.3 for an explanation on $\hat{\tau}_{high}$ and $\hat{\tau}_{low}$.

5.2.2 Math grading in Brazil

Since we used a regression forest for the detection of heterogeneity for the data set of Botelho et al. (2015), we can not calculate the average treatment effect since there is no treatment involved in a regression forest. However, we can calculate the average predictions of the dependent variable in the forest. In our case the dependent variable is the z-score of the math test in 2010. We got an average prediction of -0.031 given all the features involved in the regression forest.

Additionally, we could not apply the hypothesis test for heterogeneity using the average treatment effect, but we used the average predictions of the dependent variable. These results are shown in table 8 along with the results of the calibration test. We observe a p-value smaller than 0.05 for both tests, which specifies to reject the null hypothesis of no heterogeneity.

Table 8: The results of the hypothesis tests for heterogeneity.

	Estimate	S.E.	p-value	t-value	p-value
Calibration test	1.020	0.003	0.000***	-	-
t-test	-	-	-	346.29	0.000***

Note: The significance code *** stands for a strong significance.

The most important features that contribute to the heterogeneity are shown in table 9. We observe that high math grades in 2010 are the most important in the detection of heterogeneity. Along the fact that also the past math grade in 2009 has a importance of 16.8% in the detection of heterogeneity and the ability of reading interacted with the past math grades.

Table 9: The importance of the five most important variables for heterogeneity.

Variable	abvmedMAT2010	proficMAT2010	pastMATgrade	interpresarespMATs	interzscore_mat2010_sq
Importance	27.8%	19.2%	16.8%	6.6%	6.1%

Note: The variable names are the ones given in the data set. abvmedMAT2010 refers to 1 if the student has a math grade above the median. proficMAT2010 has a value of 1 if the student passed the math examination. pastMATgrade refers to the standardized math grade of 2009. interpresarespMATs corresonds to the squared of the past reading of the student interacted with the past math grade. interzscore_mat2010_sq is also an interaction value interacting the z-scores of the reading test with the squared value of the z-scores of the math test.

As shown in table 10, the top five most important variables for the identification of heterogeneity all have a p-value smaller than 0.05. Furthermore, we observe for higher math grades in 2009, 2010 and the interaction of the high reading ability a positive coefficient with regards to lower math grades and a lower ability for reading. The sign of the sample used for values lower than the median is for all the five variables negative as shown in the column for $\hat{\tau}_{low}$. For example, the past math grade in 2009 (pastMATgrade) has for the sample size with grades higher than the median a coefficient of 0.451 on the z-scores of the math grades in 2010, while the sample size with grades lower than the median has a negative effect of -0.507 on the dependent variable.

Table 10: Heterogeneous treatment effects of the most important variables.

Variables	$\hat{\tau}_{high}$	$\hat{\tau}_{low}$	t-value	p-value
abvmedMAT2010	0.969	-0.549	373.130	0.000
	(0.003)	(0.003)		
proficMAT2010	0.171	-1.718	484.960	0.000
	(0.003)	(0.002)		
pastMATgrade	0.451	-0.507	148.110	0.000
	(0.004)	(0.004)		
interpresarespMATs	0.336	-0.397	104.200	0.000
	(0.005)	(0.005)		
interzscore_mat2010_sq	0.323	-0.385	99.921	0.000
	(0.005)	(0.004)		

Note: The corresponding standard error is shown in brackets and for an explanation on the variables we refer to the tablenotes of table 9.

5.2.3 BRIGHT schools in Burkina Faso

When implementing the Random Forests algorithm on the data set of the BRIGHT schools in Burkina Faso, we obtained an average treatment effect of 0.048 with a corresponding standard error of 0.023 using the normalized math test score as the dependent variable. As shown in table 11, we can reject the null hypothesis of no heterogeneity for both the calibration test and the t-test indicating that there is indeed heterogeneity in the treatment effects of the BRIGHT schools in Burkina Faso.

Table 11: The results of the test for heterogeneity.

	Estimate	S.E.	p-value	t-value	p-value
Calibration test	1.484	0.646	0.011*	-	-
t-test	-	-	-	98.827	0.000***

Note: The significance codes *** stand for a strong significance, while * stand for a not as strong significance.

Table 12 presents the five most important variables for the heterogeneity in the treatment effect. We see that the amount of household members has the highest importance of 8.0%. Also other test scores and the number of kids in each household contribute to the heterogeneity of the treatment effect.

Table 12: The importance of the five most important variables for heterogeneity.

Variable	Amount of members in household	Total test score normalized	Counting test score	Number of kids in household	Counting test score normalized
Importance	8.0%	7.5%	6.3%	5.7%	5.4%

For all the five most important variables we can see in table 13 a p-value smaller than 0.05 implying we can again reject the null hypothesis of no importance for heterogeneity coming from the corresponding variables. We can also see for the variables containing the test scores that for the observations containing values higher than the median of the corresponding test scores have a higher average treatment effect than the sample containing observations lower than the median. For example, the normalized total test score has a $\hat{\tau}_{high}$ of 0.098, while $\hat{\tau}_{low}$ is 0.002.

Table 13: Heterogeneous treatment effects of the most important variables.

Variables	$\hat{\tau}_{high}$	$\hat{\tau}_{low}$	t-value	p-value
Amount of members in household	0.055	0.054	10.027	0.000
	(0.001)	(0.000)		
Total test score normalized	0.098	0.002	70.853	0.000
	(0.001)	(0.000)		
Counting test score	0.103	0.028	51.694	0.000
	(0.001)	(0.000)		
Number of kids in houehold	0.056	0.052	-3.870	0.000
	(0.001)	(0.000)		
Counting test score normalized	0.103	0.028	51.694	0.000
	(0.001)	(0.000)		

Note: The corresponding standard error is shown in brackets and for an explanation on the variables we refer to the tablenotes of table 9.

6 Conclusion

The focus of our research was on the implementation of the Random Forests algorithm to answer our main research question: *"Can we find other heterogeneous treatment effects that have not*

been investigated before by previous researchers using the Random Forests algorithm? And if so, which ones?" in three different contexts.

Our first instance was in the context of the Romanian Baccalaureate scores, where we first replicated the results of the work of Borcan et al. (2017). We fully obtained the same estimates for each case, whereas our conclusion for these results also equals the conclusion in the paper of Borcan et al. (2017). Out of this analysis, we can conclude that the campaign against corruption did reduce the fraud among the high-school students. However, we also found heterogeneity in the students' poverty status saying that the poor high-school students performed even worse after the launching of the campaign in 2011. This result was not in line with the hypothesis of the authors, but is really important for policy makers. Policy makers can use this result to focus the campaign more on students with a poor background and try to better inform them on the consequences of committing fraud during the Baccalaureate.

When implementing the Random Forests algorithm, we found out that when using the overall examination score as dependent variable in the Random Forests algorithm, the algorithm successfully detects heterogeneous treatment effects. We can also conclude that one of the most important variables for detecting heterogeneity is the percentile ranking of the written examination, but also the mid school GPA and the ability are in the top five most important variables. Since, the midschool GPA is used in the data set as a control variable for ability, it is expected that both of these variables are of importance. Thus, to come back to our research question we successfully found other heterogeneous treatment effects than Borcan et al. (2017), because they only found heterogeneity with regards to the students' poverty status, but with the results in table 6 we can conclude that there are more heterogeneous treatment effects such as the ability of a student.

Moreover, when calculating the heterogeneous treatment effect for the variable percentile ranking of the written examination we can conclude that the students that are ranked low in this percentile ranking have a much larger response in the outcome of the overall examination score. This is also the same for the mid school GPA and the ability meaning the students with a lower mid school GPA and a lower ability have a much lower score when monitored by a CCTV camera than students with a higher mid school GPA and a higher ability. We did the same approach for the variable poor to determine if we could also say that there is heterogeneity with regards to the poverty status like Borcan et al. (2017) found in their work. We indeed can reject the null hypothesis of the t-test implying that the variable poor has a impact on the heterogeneity. However, this variable is not as important as the variables found in table 6.

Our second application was in the context of math grading in Brazil. In the paper of Botelho et al. (2015), they only looked into heterogeneous effects of race on the math grades of 2010 in Brazil. However, we found out with the Random Forests algorithm that there are more heterogeneous effects other than just the race of the students. In table 9 we can conclude that also the past math grade of 2009 and the interaction with the ability reading have a strong impact on the identification of heterogeneity. It is remarkable again for this application that the variable where past researchers have focused on, in this case the race of a student, did not appear in our top five most important variables.

Additionally, out of the results shown in table 10 we can conclude that the top five most

important variables all have a significant effect in the detection of heterogeneity. We can also say that the students with a higher past math grades and a higher ability for reading, will more likely have a higher z-score for the math grade in 2010.

The last application on the Random Forests algorithm was in the context of BRIGHT schools in Burkina Faso. Kazianga et al. (2012) did not look into any heterogeneous treatment effects. So, to answer our main research question we found out that the number of members and kids in a household have a heterogeneous treatment effect on the math grades of primary-school students. In table 13 we can conclude that students from households with more members and more kids will have a slightly higher normalized math score when attending a BRIGHT school than students from households with less members and less kids. Also, kids with higher test scores for counting numbers have a larger chance to have a higher normalized math score when attending a BRIGHT school than kids with lower test scores for counting numbers. We successfully found new heterogeneous treatment effects in this context.

Policy makers of Romania can use these results and conclusions obtained from the Random Forests algorithm to focus their campaign on students with a lower ability than on students with a poor background, since we found out that variables concerning ability have a much bigger importance on the heterogeneous treatment effects than the variable poverty status. For economists in Brazil studying about racial discrimination in education now know with this result that the variation of math grades in 2010 did not only rely on the racial differences, but are also due to past math grades and the reading ability of the students. Teachers can use this result to improve the reading ability of students, such that they also improve their math ability. And lastly founders of the BRIGHT schools in Burkina Faso can use these results for further improvement on the BRIGHT schools. They know now that the more members and kids in a household the higher the normalized math scores are for kids attending a BRIGHT school and they can improve the amenities of a BRIGHT schools such that the kids from households with less members and children will have the same effect for the normalized math scores as the kids from households with more members and children.

Further research on this topic need to focus on the improvement of the **grf** package used in this thesis implemented by Athey et al. (2019). Even though, the Random Forests algorithm is in the literature known for handling missing values quite well. The **grf** package is not resistant for these missing values and will only provide predictions when the data used for the algorithm does not contain any missing values.

Additionally, the **grf** package is quite new in **R**, which means that the authors of the code still publish new versions every few months. The last version just published on the 27th of May. This implies that the code will still contain flaws and improvements that will be fixed when time goes on. For example, printing a tree in the forest does not give a clear image on the splitting variables and splitting values. This is especially the case if you use a lot of variables in the algorithm. If the code could provide the splitting values in a more clear way, it would be better to use this value for the criteria in splitting the sample in two groups for calculating the heterogeneous treatment effects. Now we use the median to provide these two groups, but this is obviously not the real value, where the tree split on.

References

A. Abadie, S. Athey, G. Imbens, and J. Wooldridge. When should you adjust standard errors for clustering? 10 2017.

C. Ai and E. Norton. Interaction terms in logit and probit models. *Economics Letters*, 80:123–129, 07 2003. doi: 10.1016/S0165-1765(03)00032-6.

O. Ashenfelter and D. Card. Using the longitudinal structure of earnings to estimate the effect of training programs. (1489), November 1984. doi: 10.3386/w1489.

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113:7353–7360, 07 2016. doi: 10.1073/pnas.1510489113.

S. Athey and S. Wager. Estimating treatment effects with causal forests: An application. 02 2019.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Ann. Statist.*, 47(2):1148–1178, 04 2019. doi: 10.1214/18-AOS1709. URL https://doi.org/10.1214/18-AOS1709.

O. Borcan, M. Lindahl, and A. Mitrut. Fighting corruption in education: What works and who benefits? *American Economic Journal: Economic Policy*, 9(1):180–209, February 2017. doi: 10.1257/pol.20150074. URL http://www.aeaweb.org/articles?id=10.1257/pol.20150074.

F. Botelho, R. A. Madeira, and M. A. Rangel. Racial discrimination in grading: Evidence from brazil. *American Economic Journal: Applied Economics*, 7(4):37–52, October 2015. doi: 10.1257/app.20140352. URL http://www.aeaweb.org/articles?id=10.1257/app.20140352.

L. Breiman. Out-of-bag estimation. 12 1996.

L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, Oct. 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. (1712.04802), Dec. 2017. URL https://ideas.repec.org/p/arx/papers/1712.04802.html.

T. Conley and C. Taber. Inference with "difference in differences" with a small number of policy changes. (312), July 2005. doi: 10.3386/t0312. URL http://www.nber.org/papers/t0312.

M. Gebel and J. Voßemer. The impact of employment transitions on health in germany. a difference-in-differences propensity score matching approach. *Social Science Medicine*, 108:128–136, 05 2014. doi: 10.1016/j.socscimed.2014.02.039.

H. Kazianga, D. Levy, L. Linden, and M. Sloan. The effects of girl-friendly schools: Evidence from the bright school construction program in burkina faso. *American Economic Journal Applied Economics*, 5, 05 2012. doi: 10.1257/app.5.3.41.

M. Knaus, M. Lechner, and A. Strittmatter. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. 10 2018.

N. Korepanova, H. Seibold, V. Steffen, and T. Hothorn. Survival forests under test: Impact of the proportional hazards assumption on prognostic and predictive forests for als survival. 02 2019.

M. Lechner. The estimation of causal effects by difference-in-difference methods. 2011. URL https://EconPapers.repec.org/RePEc:usg:dp2010:2010-28.

M. Oprescu, V. Syrgkanis, and Z. S. Wu. Orthogonal random forest for heterogeneous treatment effect estimation. *CoRR*, abs/1806.03467, 2018.

E. Scornet. On the asymptotics of random forests. *J. Multivariate Analysis*, 146:72–83, 2016.

D. Viviano and J. Bradic. Synthetic learner: model-free inference on treatments over time. 04 2019.

S. Wager. Asymptotic theory for random forests. 05 2014.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

G. Wang, J. li, and W. Hopp. An instrumental variable tree approach for detecting heterogeneous treatment effects in observational studies. *SSRN Electronic Journal*, 01 2017. doi: 10.2139/ssrn.3045327.

7 Appendix

A: Zip-file

Table 14: Zip-file files description.

File name	Description
DID_romania.do	Generates tables 1 to 3 with the results of the DID-estimation
data1.dta	The data set of the paper of Borcan et al. (2017)
forest1_romania.R	Generates tables 4 to 7 with the results of the Random Forests algorithm applied on data1.dta
data2.xlsx	The data set of the paper of Botelho et al. (2015)
forest2_brazil.R	Generates tables 8 to 10 with the results of the Random Forests algorithm applied on data2.xlsx
data3.dta	The data set of the paper of Kazianga et al. (2012)
forest3_burkinafaso.R	Generates tables 11 to 13 with the results of the Random Forests algorithm applied on data3.dta

B: Random Forests code

B.1: Romanian Baccalaureat scores

```
library(grf)
library(readxl)
library(foreign)

#Get the data
dataset <- read.dta("Desktop/THESIS_ZIP/data1.dta")
data <- na.omit(dataset)

student_id <- data[,1]

count = 0
for(i in student_id){
  if (i <= 800000){
    count = count + 1
  }
}

X = matrix(0, count, 44)

index = 1
count1 = 0
for (j in student_id){
  count1 = count1 + 1
  if (j <= 800000){ #work with student-id smaller than 800000
    for (k in 1:ncol(data)){
```

```
      X[index,k] <- data[count1,k]
    }
    index = index + 1
  }
}

W = X[,20] #new_cam

#Train a causal forest with outcome: overall examination score (1)
c.forest1 = causal_forest(X,X[,30],W, clusters = X[,3], seed = 7)
c.pred1 = c.forest1$predictions

#Train a causal forest with outcome: examination pass (2)
c.forest2 = causal_forest(X,X[,10],W, clusters = X[,3], seed = 7)
c.pred2 = c.forest2$predictions

#Train a causal forest with outcome: written examination score (3)
c.forest3 = causal_forest(X,X[,7],W, clusters = X[,3], seed = 7)
c.pred3 = c.forest3$predictions

#Table 4: Average treatment effect estimates per outcome.
ate1 = average_treatment_effect(c.forest1, target.sample = "overlap")
ate2 = average_treatment_effect(c.forest2, target.sample = "overlap")
ate3 = average_treatment_effect(c.forest3, target.sample = "overlap")
ci_se1 = c( round(ate1,3) - round(qnorm(0.975)*ate1[2], 3), round(ate1,3) +
    round(qnorm(0.975)*ate1[2], 3))
ci1 = c(ci_se1[3], ci_se1[1])
ci_se2 = c( round(ate2,3) - round(qnorm(0.975)*ate2[2], 3), round(ate2,3) +
    round(qnorm(0.975)*ate2[2], 3))
ci2 = c(ci_se2[3], ci_se2[1])
ci_se3 = c( round(ate3,3) - round(qnorm(0.975)*ate3[2], 3), round(ate3,3) +
    round(qnorm(0.975)*ate3[2], 3))
ci3 = c(ci_se3[3], ci_se3[1])

#Table 5: the results of the hypothesis tests for heterogeneity per outcome.
#Test 1: calibration test
test_calibration(c.forest1)
test_calibration(c.forest2)
test_calibration(c.forest3)
#Test 2: t-test
high_effect1 = c.pred1 > median(c.pred1)
ate_high1 = average_treatment_effect(c.forest1, subset = high_effect1[1:47909])
ate_low1 = average_treatment_effect(c.forest1, subset = !high_effect1[1:47909])
ate_diff1 = ate_high1 - ate_low1
ci_hl1 = c( round(ate_diff1, 3) - round(qnorm(0.975)*sqrt(ate_high1[2]^2 +
    ate_low1[2]^2), 3),
           round(ate_diff1, 3) + round(qnorm(0.975)*sqrt(ate_high1[2]^2 +
               ate_low1[2]^2), 3))
```

```
t1 = t.test(c.pred1[high_effect1], c.pred1[!high_effect1])

high_effect2 = c.pred2 > median(c.pred2)
ate_high2 = average_treatment_effect(c.forest2, subset = high_effect2[1:47909])
ate_low2 = average_treatment_effect(c.forest2, subset = !high_effect2[1:47909])
ate_diff2 = ate_high2 - ate_low2
ci_hl2 = c( round(ate_diff2, 3) - round(qnorm(0.975)*sqrt(ate_high2[2]^2 +
    ate_low2[2]^2), 3),
           round(ate_diff2, 3) + round(qnorm(0.975)*sqrt(ate_high2[2]^2 +
               ate_low2[2]^2), 3))
t2 = t.test(c.pred2[high_effect2], c.pred2[!high_effect2])

high_effect3 = c.pred3 > median(c.pred3)
ate_high3 = average_treatment_effect(c.forest3, subset = high_effect3[1:47909])
ate_low3 = average_treatment_effect(c.forest3, subset = !high_effect3[1:47909])
ate_diff3 = ate_high3 - ate_low3
ci_hl3 = c( round(ate_diff3, 3) - round(qnorm(0.975)*sqrt(ate_high3[2]^2 +
    ate_low3[2]^2), 3),
           round(ate_diff3, 3) + round(qnorm(0.975)*sqrt(ate_high3[2]^2 +
               ate_low3[2]^2), 3))
t3 = t.test(c.pred3[high_effect3], c.pred3[!high_effect3])

#Table 6: Variable Importance per outcome (order from high to low)
variableimp1 = variable_importance(c.forest1)
variableimp2 = variable_importance(c.forest2)
variableimp3 = variable_importance(c.forest3)

#Table 7: Heterogeneous treatment effects of the most important variables
ranking_wr = X[,28] > median(X[,28]) #percentile ranking written examination score
ranking_ate_high = average_treatment_effect(c.forest1, subset = ranking_wr[1:47909])
ranking_ate_low = average_treatment_effect(c.forest1, subset = !ranking_wr[1:47909])
ranking_diff = ranking_ate_high - ranking_ate_low
ranking_t = t.test(c.pred1[ranking_wr], c.pred1[!ranking_wr])

midschool_gpa = X[,17] > median(X[,17]) #midschool gpa
midschool_ate_high = average_treatment_effect(c.forest1, subset =
    midschool_gpa[1:47909])
midschool_ate_low = average_treatment_effect(c.forest1, subset =
    !midschool_gpa[1:47909])
midschool_diff = midschool_ate_high - midschool_ate_low
midschool_t = t.test(c.pred1[midschool_gpa], c.pred1[!midschool_gpa])

ability = X[,33] > median(X[,33]) #ability
ability_ate_high = average_treatment_effect(c.forest1, subset = ability[1:47909])
ability_ate_low = average_treatment_effect(c.forest1, subset = !ability[1:47909])
ability_diff = ability_ate_high - ability_ate_low
ability_t = t.test(c.pred1[ability], c.pred1[!ability])
```

```
poor = X[,16] > median(X[,16]) #poor
poor_ate_high = average_treatment_effect(c.forest1, subset = poor[1:47909])
poor_ate_low = average_treatment_effect(c.forest1, subset = !poor[1:47909])
poor_diff = poor_ate_high - poor_ate_low
poor_t = t.test(c.pred1[poor], c.pred1[!poor])
```

B.2: Math grading in Brazil

```
library(readxl)
library(foreign)
library(grf)
library(plotrix)

extra <- read_excel("Desktop/THESIS_ZIP/data2.xlsx")
data <- na.omit(extra)
X = data[1:50000,]
Z = X[,-c(4)]
Y = X[,4]

#Train a regression forest
r.forest = regression_forest(X[,-c(4)], X[,4], seed = 7)
r.prediction = predict(r.forest)
r.pred = r.prediction$predictions
mean(r.pred)

#Table 8: the results of the hypothesis tests for heterogeneity
test_calibration(r.forest)
avg = mean(r.pred)
high_effect = r.pred > mean(r.pred)
t_test = t.test(r.pred[high_effect], r.pred[!high_effect])

#Table 9: variable importance
varimp = variable_importance(r.forest)

#Table 10: heterogeneous treatment effects of the most important variables
abvmed = Z[,5] == 100 #abvmedMAT2010
abvmed_high_mean = mean(r.pred[abvmed])
abvmed_low_mean = mean(r.pred[!abvmed])
abvmed_high_se = std.error(r.pred[abvmed])
abvmed_low_se = std.error(r.pred[!abvmed])
abvmed_t = t.test(r.pred[abvmed], r.pred[!abvmed])

profic = Z[,4] == 100 #proficMAT2010
profic_high_mean = mean(r.pred[profic])
profic_low_mean = mean(r.pred[!profic])
profic_high_se = std.error(r.pred[profic])
profic_low_se = std.error(r.pred[!profic])
```

```
profic_t = t.test(r.pred[profic], r.pred[!profic])

past = Z[,33] > median(Z[,33]) #pastMATgrade
past_high_mean = mean(r.pred[past])
past_low_mean = mean(r.pred[!past])
past_high_se = std.error(r.pred[past])
past_low_se = std.error(r.pred[!past])
past_t = t.test(r.pred[past], r.pred[!past])

inter = Z[,48] > median(Z[,48]) #interpresarespMATs
inter_high_mean = mean(r.pred[inter])
inter_low_mean = mean(r.pred[!inter])
inter_high_se = std.error(r.pred[inter])
inter_low_se = std.error(r.pred[!inter])
inter_t = t.test(r.pred[inter], r.pred[!inter])

interz = Z[,39] > median(Z[,39]) #interzscore_mat2010_sq
interz_high_mean = mean(r.pred[interz])
interz_low_mean = mean(r.pred[!interz])
interz_high_se = std.error(r.pred[interz])
interz_low_se = std.error(r.pred[!interz])
interz_t = t.test(r.pred[interz], r.pred[!interz])
```

B.3: BRIGHT schools in Burkina Faso

```
library(grf)
library(haven)
maindata <- read_dta("Desktop/THESIS_ZIP/data3.dta")
maindata1 <-
    maindata[,-c(43,44,45,46,47,49,60,66,68,69,73,74,82,83,84,85,86,87,88,89,90,91,106,107,111)]
omit <- na.omit(maindata1)
X <- data.matrix(omit, rownames.force = NA)

Y = X['math_norm']
W = X['MCC_School']

#Train a causal forest
c.forest = causal_forest(X ,Y , W ,seed = 7)
ate = average_treatment_effect(c.forest, target.sample = "overlap")
c.pred = c.forest$predictions

#Table 11: the results of the test for heterogeneity
test_calibration(c.forest)
high_effect = c.pred > median(c.pred)
ate_high = average_treatment_effect(c.forest, subset = high_effect[1:17664])
ate_low = average_treatment_effect(c.forest, subset = !high_effect[1:17664])
ate_diff = ate_high - ate_low
```

```
ci_hl = c( round(ate_diff, 3) - round(qnorm(0.975)*sqrt(ate_high[2]^2 +
    ate_low[2]^2), 3),
          round(ate_diff, 3) + round(qnorm(0.975)*sqrt(ate_high[2]^2 + ate_low[2]^2),
              3))
t = t.test(c.pred[high_effect], c.pred[!high_effect])

#Table 12: the variable importance
variableimp = variable_importance(c.forest)

#Table 13: Heterogeneous treatment effects
numMem_wr = X[,36] > median(X[,36]) #Number members of household
numMem_ate_high = average_treatment_effect(c.forest, subset = numMem_wr[1:17664])
numMem_ate_low = average_treatment_effect(c.forest, subset = !numMem_wr[1:17664])
numMem_diff = numMem_ate_high - numMem_ate_low
numMem_t = t.test(c.pred[numMem_wr], c.pred[!numMem_wr])

totNorm_gpa = X[,84] > median(X[,84]) #total normalized score
totNorm_ate_high = average_treatment_effect(c.forest, subset = totNorm_gpa[1:17664])
totNorm_ate_low = average_treatment_effect(c.forest, subset = !totNorm_gpa[1:17664])
totNorm_diff = totNorm_ate_high - totNorm_ate_low
totNorm_t = t.test(c.pred[totNorm_gpa], c.pred[!totNorm_gpa])

countRaw = X[,58] > median(X[,58]) #counting raw score
countRaw_ate_high = average_treatment_effect(c.forest, subset = countRaw[1:17664])
countRaw_ate_low = average_treatment_effect(c.forest, subset = !countRaw[1:17664])
countRaw_diff = countRaw_ate_high - countRaw_ate_low
countRaw_t = t.test(c.pred[countRaw], c.pred[!countRaw])

numKids = X[,35] > median(X[,35]) #number of kids in household
numKids_ate_high = average_treatment_effect(c.forest, subset = numKids[1:17664])
numKids_ate_low = average_treatment_effect(c.forest, subset = !numKids[1:17664])
numKids_diff = numKids_ate_high - numKids_ate_low
numKids_t = t.test(c.pred[numKids], c.pred[!numKids])

countNorm = X[,57] > median(X[,57]) #counting normalized score
countNorm_ate_high = average_treatment_effect(c.forest, subset = countNorm[1:17664])
countNorm_ate_low = average_treatment_effect(c.forest, subset = !countNorm[1:17664])
countNorm_diff = countNorm_ate_high - countNorm_ate_low
countNorm_t = t.test(c.pred[countNorm], c.pred[!countNorm])
```

C: Elaborate tables for the DID-estimation

Table 15: Main results of the DID estimation.

	Written Examination Score				Examination Pass				Overall Examination Score			
Camera		-0.246	-0.251	-0.353		-0.076	-0.076	-0.095		-0.430	-0.439	-0.512
		(0.108)	(0.106)	(0.106)		(0.03)	(0.029)	(0.025)		(0.144)	(0.142)	(0.137)
2012	-0.874	-0.628	-0.716	-0.463	-0.211	-0.135	-0.148	-0.082	-0.923	-0.492	-0.579	-0.323
	(0.065)	(0.087)	(0.078)	(0.087)	(0.024)	(0.025)	(0.023)	(0.017)	(0.092)	(0.115)	(0.106)	(0.094)
2011	-0.875	-0.713	-0.743	-0.597	-0.211	-0.161	-0.166	-0.129	-0.943	-0.660	-0.690	-0.547
	(0.058)	(0.07)	(0.071)	(0.081)	(0.022)	(0.019)	(0.018)	(0.016)	(0.088)	(0.091)	(0.09)	(0.088)
2009	-0.205	-0.205	-0.237	-0.311	0.121	0.121	0.115	0.093	1.087	1.086	1.055	0.967
	(0.054)	(0.054)	(0.057)	(0.033)	(0.011)	(0.011)	(0.011)	(0.012)	(0.042)	(0.042)	(0.04)	(0.04)
Male			-0.852	-0.852			-0.109	-0.109			-0.590	-0.590
			(0.016)	(0.016)			(0.003)	(0.003)			(0.014)	(0.013)
Poor			-0.224	-0.222			-0.045	-0.044			-0.263	-0.260
			(0.022)	(0.022)			(0.004)	(0.004)			(0.02)	(0.02)
Track			1.458	1.457			0.318	0.318			1.559	1.559
			(0.049)	(0.049)			(0.013)	(0.012)			(0.052)	(0.051)
Rural			-0.662	-0.666			-0.137	-0.137			-0.652	-0.656
			(0.02)	(0.02)			(0.02)	(0.02)			(0.085)	(0.085)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
County FE × Yearly trends	No	No	No	Yes	No	No	No	Yes	No	No	No	Yes
Observations	712,298	712,298	712,298	712,298	731,505	731,505	731,505	731,505	706,895	706,895	706,895	706,895
R^2	0.06	0.06	0.275	0.289	0.102	0.103	0.239	0.253	0.204	0.206	0.417	0.432

Note: the corresponding standard error is shown in brackets.

Table 16: Heterogeneous effects by using the poverty status of students

	Written Examination Score			Examination Pass			Overall Examination Score		
	1	2	3	4	5	6	7	8	9
Camera	-0.302	-0.255	-0.131	-0.081	-0.088	-0.077	-0.433	-0.434	-0.308
	(0.11)	(0.082)	(0.056)	(0.026)	(0.023)	(0.017)	(0.141)	(0.113)	(0.076)
Poor × Camera	-0.213	-0.257	-0.214	-0.062	-0.056	-0.051	-0.346	-0.350	-0.306
	(0.063)	(0.053)	(0.052)	(0.015)	(0.012)	(0.013)	(0.078)	(0.061)	(0.063)
Year fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County fixed effect × Yearly trends	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Poor interactions	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ability interactions	No	No	Yes	No	No	Yes	No	No	Yes
Observations	712,298	547,447	547,447	731,505	553,903	553,903	706,895	545,121	545,121
R^2	0.291	0.356	0.459	0.256	0.31	0.394	0.435	0.504	0.613

Note: the corresponding standard error is shown in brackets.