Heterogeneous Treatment Effects in Regression Discontinuity Design

Ágoston Reguly*

Central European University

October 21, 2021

Abstract

The paper proposes a causal supervised machine learning algorithm to uncover treatment effect heterogeneity in classical regression discontinuity (RD) designs. Extending Athey and Imbens (2016), I develop a criterion for building an honest "regression discontinuity tree", where each leaf of the tree contains the RD estimate of a treatment (assigned by a common cutoff rule) conditional on the values of some pre-treatment covariates. It is *a priori* unknown which covariates are relevant for capturing treatment effect heterogeneity, and it is the task of the algorithm to discover them, without invalidating inference. I study the performance of the method through Monte Carlo simulations and apply it to the data set compiled by Pop-Eleches and Urquiola (2013) to uncover various sources of heterogeneity in the impact of attending a better secondary school in Romania.

JEL: C13, C21, I21

Keywords: Supervised machine learning, regression tree, regression discontinuity design, heterogeneous treatment effect, CATE.

*I am grateful for all the guidance and the thorough comments by my advisor Róbert Lieli, without him this paper would not exist in this form. I thank two of my Ph.D. examiners Gábor Békés and Michael Knaus for their helpful comments and suggestions. I also thank to participants of EEA-ESEM 2021 and IAAE 2021 conferences, Lajos Szabó, Arieda Muço, László Mátyás, Balázs Vonnák, János Divényi, Andrea Weber, Sergey Lychagin, Tímea Molnár and for the participants at the Brownbag Seminar Series at CEU for helpful comments and suggestions. The usual disclaimer applies.

Email address: reguly_agoston@phd.ceu.edu

Codes are available at `https://github.com/regulyagoston/RD_tree`

1 Introduction

In regression discontinuity (RD) designs one identifies the *average* treatment effect from a jump in the regression function caused by the change in treatment assignment (or the probability of treatment assignment) as a running variable crosses a given threshold. Identification is based on comparing outcomes on the two sides of the cutoff, assuming that all other factors affecting the outcome change continuously with the running variable, which is not manipulable (see, e.g., Hahn et al., 2001, Imbens and Lemieux (2008), Lee and Lemieux (2010), Calonico et al., 2014). From 2005 regression discontinuity has become extremely popular in theoretical and empirical works, resulting in a large number of extensions.[1]

This paper contributes to the literature by proposing a machine learning algorithm designed to discover heterogeneity in the average treatment effect (ATE) estimated in an RD setup. The subpopulations that the algorithm searches over are defined by the values of a set of additional pre-treatment covariates. Analysis of treatment effect heterogeneity is important for at least two reasons. Firstly, researchers and policymakers gain a more detailed understanding of the treatment by learning the extent to which the treatment works differently in different groups. Indeed, the overall average effect may not be very informative if there is substantial heterogeneity. For example, the treatment may have no impact in one group while a large one in another, or there may even be groups where the average treatment effect has opposite signs. Secondly, uncovering treatment effect heterogeneity – with strong external validity – can lead to a more efficient allocation of resources. If the budget for implementing a treatment is limited, decision makers can design future policies to focus on treating those groups where the expected treatment effects are the largest.

Of course, heterogeneity analysis is routinely undertaken in applied work, typically by repeating the main RD estimation within different groups defined by the researcher. Nevertheless, ad-hoc (or even pre-specified) selection of sub-samples has disadvantages: i) when there are many candidate groups defined by pre-treatment covariates, searching across these groups presents a multiple testing problem and without correction, it leads to invalid inference. ii) The relevant groups may have a complicated non-linear relationship with the treatment effect and discovering the non-linear pattern is cumbersome or impossible "by

[1] E.g., Becker et al. (2013) defines heterogeneous local treatment effects in RD, where heterogeneity comes from a known covariate; Calonico et al. (2019) analyze the effect of using additional covariates; Xu (2017) extend the analysis with categorical variables as outcome; Cattaneo et al. (2016) concern multiple thresholds; Caetano et al. (2017) uses covariates to generate over-identifying restrictions in case of multiple treatment variable; Robson et al. (2019) proposes decomposition of ATE and CATE using covariate(s) with non-parametric methods; Toda et al. (2019) uses multiple groups with multiple threshold values to estimate CATE given by these pre-specified groups; Toda et al. (2019) uses machine learning to find discontinuity when there are many (potential) running variables and thresholds – but no heterogeneity in the treatment effect. Cattaneo et al. (2019) gives a great overview of recent developments in RD.

hand." For example, searching along the interactions of the pre-treatment covariates is usually infeasible and the researcher only checks few interactions motivated by theoretical considerations.[2]

By contrast, the method proposed in this paper allows discovering treatment effect heterogeneity systematically based on pre-treatment covariates while offering a solution to the aforementioned challenges. At present, I know of no other paper that accomplishes these goals specifically in an RD setup. The closest paper with an RD focus is perhaps Hsu and Shen (2019), which develop *tests* for possible heterogeneity in the treatment effect based on the null hypothesis that the conditional average treatment effect (CATE) function is equal to a constant (the overall average treatment effect). Their proposed tests reveal whether there are groups defined in terms of observed characteristics for which the ATE deviates from the overall average, but they leave the discovery and the estimation of the conditional average treatment effect function as an open question. I address precisely this problem by proposing a data-driven machine learning method, which discovers groups with different treatment effects, using many candidate pre-treatment variables, without invalidating inference. The method provides discovery in the sense that the researcher does not need to specify the sources of heterogeneity (the relevant variables) in a pre-analysis plan, but can use many potentially relevant pre-treatment variables. The task of the algorithm is to find the relevant variables and the functional form from the many possible combinations. The end result gives groups with differences in the treatment effects. The implementation of the algorithm assumes that the standard RD identification conditions hold in the potentially relevant subpopulations; e.g., one cannot consider groups in which the running variable is always above or below the cutoff.

The paper also builds on and extends the more recent literature of discovering heterogeneous treatment effects with machine learning methods. There is a growing number of papers (e.g., Imai et al. (2013), Athey and Imbens (2016), Wager and Athey (2018), Athey et al. (2019) Bargagli and Gnecco (2020), Friedberg et al. (2020), Knaus (2021) or Knaus et al. (2021)) using causal supervised machine learning (ML) techniques for this purpose.[3] All of these works are concerned with i) randomized experiments and/or ii) observational studies with the unconfoundedness assumption or iii) using instruments to estimate the local average treatment effect (LATE). Imai et al. (2013) use lasso with two sparsity constraints to

[2] Hsu and Shen (2019) carry out a small survey of top publications in economics in 2005 that use the RD design. They find that 15 out of 17 papers check for heterogeneity, and only 2 address the issue with interaction terms. The rest use subsample techniques without correcting for multiple testing.

[3] There is another, distinct, strand of the broader causal inference literature where ML techniques are used for estimating high-dimensional nuisance parameters, while the parameter of interest is still the average treatment effect or a reduced dimensional version of CATE. See e.g., Chernozhukov et al. (2018), Semenova and Chernozhukov (2020) or Fan et al. (2020).

identify heterogeneous treatment effects. The idea is to formulate heterogeneity as a variable selection problem in randomized experiments or observational studies with the unconfoundedness assumption. Athey and Imbens (2016) also focus on randomized experiments or observational studies with unconfoundedness, but use what they call *honest* regression trees to find heterogeneity in the treatment effect. The honest approach means that independent samples are used for growing the tree and estimating the average treatment effect in the resulting leaves. This ensures that traditional confidence intervals constructed for the estimates have the proper coverage rate. Bargagli and Gnecco (2020) follow the Athey and Imbens (2016) approach and extend it with instrumental variable setting to estimate conditional local average treatment effects (CLATE). Finally, the rest of the aforementioned papers and references therein go beyond regression trees[4] and use random forests or other machine learning methods to estimate conditional treatment effects in settings i), ii) and iii).[5] As the aforementioned methods are already developed, one may argue that CATE function in RD can be estimated by using these causal supervised machine learning methods. However, these methods require some restrictive and unnecessary assumptions when identifying the ATE parameter in RD context. For example, in observational studies with unconfoundedness, it is impossible to construct a treatment-control contrast, without overlap. The lack of overlap causes the propensity score to take either the value of 1 or 0. This is indeed a problem as in many causal supervised machine learning methods this is typically excluded by assumption. E.g., for the propensity score weighted outcomes to make the transformation possible, one needs to assume that the propensity score values are away from the boundaries. Another caveat is that these methods assume the unobservable factors must be the same for treated and control units for every value of the running variable, thus the treated and control conditional expectation functions (CEFs) are on top of each other. In contrast, RD only assumes continuity for CEFs around the threshold, and CEFs can be

[4]Let me note here that in this paper I discuss building only one tree, which is known to be less stable in case of variables are (highly) correlated with each other. Extension to forest methods would generate a more robust estimator from this perspective. This extension is left for future research.

[5]Wager and Athey (2018), introduces causal (random) forests and shows that using honest trees to construct the forest, yields asymptotic normality for the conditional treatment effect estimator. They implement their theoretical results for causal forests in randomized experiments or observational studies with unconfoundedness. Friedberg et al. (2020) uses 'generalized random forests' as an adaptive weighting function to express heterogeneity. Friedberg et al. (2020) improves the asymptotic rates of convergence for generalized random forests with smooth signals by using local linear regressions, where the weights are given by the forests. Their method applies to randomized experiments and shows an application with observational study with unconfoundedness assumption. Knaus (2021) synthesize different methods using double machine learning with a focus on program evaluation under unconfoundedness assumption. He also proposes a normalized DR-learner to estimate individual average treatment effects. Knaus et al. (2021) provide a great overview of the Empirical Monte Carlo Study performances of the different machine learning methods, which are available and used in practice.

anything away from the cutoff. Another approach is to use instrumental variables which relaxes the assumption of unobservable factors to be the same for both treated and control groups by using instruments. With introducing instrument(s) the core assumption is the exclusion restriction, thus instrumental variables enter only to the selection equation, but not the outcome equation, and are uncorrelated with the unobservables. Usually, it is hard to find such variable(s) in the context of RD. Furthermore, when the instrument is binary the ATE or CATE can be identified without further assumption. However, if this is not the case, one needs to use the "identification at infinity" assumption. In contrast, with classical RD design, the researcher can avoid taking such strong assumption(s) while relying on the observed running variable and the continuity assumptions.[6]

Finally, let me mention a closely related paper by Athey et al. (2019). They work out a general framework for estimating heterogeneous treatment effects called 'generalized random forests' based on local moment conditions. In their paper, they work out the local moment conditions for nonparametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation via instrumental variables. However, they do not account for regression discontinuity designs. From their perspective, this paper extends their methodology to RD designs and derives the properties of a parametric RD estimator for a single tree. [7]

I contribute to the causal machine learning literature by introducing a specialized machine learning method to search for and estimate conditional average treatment effects in an RD setup. Following Athey and Imbens (2016) I capture heterogeneity by building an honest "regression discontinuity tree", where each leaf of the tree contains a parametric RD regression (to be estimated over an independent sample) rather than a simple difference between two means.[8] Similarly, an expected mean squared error criterion used to build the tree is modified appropriately to account for the more complicated statistic to be computed within each candidate leaf. Furthermore, the tree building algorithm also needs modifications to accommodate RD estimation and the new criterion. From a strictly technical standpoint, these are the main contributions of the paper. With the proposed algorithm, one can achieve unbiased estimates for the group-level (conditional) average treatment effects and their vari-

[6]For more detailed discussion on how to estimate ATE with different types of models, see Lee and Lemieux (2010) Section 3.5.

[7]For completeness, let me mention a working paper by Nekipelov et al. (2019). They use moment-based models when constructing trees and they call it 'moment forest'. They use regression discontinuity design as an application for their method however, it is in a preliminary state: they make some strong assumptions on the functional form of conditional expectation functions when estimating the CATE function, which I do not require in this paper.

[8]My future research agenda includes allowing for nonparametric RD estimation where the search for heterogeneity and the choice of the appropriate bandwidth is handled simultaneously.

ance.

I present Monte Carlo simulations to demonstrate that the algorithm successfully discovers and estimates heterogeneity in a variety of settings — at least with suitably large samples. In addition, I use the well-known and investigated dataset of Pop-Eleches and Urquiola (2013) on the Romanian school system. Pop-Eleches and Urquiola (2013) study the average treatment effect on Baccalaureate examination outcomes of going to a better school, and undertake some additional ad-hoc heterogeneity analysis. Hsu and Shen (2019) use their proposed test and show some evidence on the heterogeneity in the treatment effect without identifying the sources of the heterogeneity. I show that using the algorithm I can refine their results, discovering important treatment heterogeneity along with the level of school average transition scores[9] and the number of schools in town. The algorithm reveals groups that have different treatment effects but were missed by Pop-Eleches and Urquiola (2013). Furthermore, with a more extensive survey dataset with many socio-economic variables (but with fewer observations), I find that the estimated intention-to-treat effect varies among other covariates with having internet access at home, gender of the student, the education of the mother, and the proportion of novice teachers in school.

The paper is organized as follows. Section 2 introduces the concept of a sharp RD, a regression tree, and defines the conditional average treatment effect for the regression discontinuity tree. Section 3 develops the honest criterion for RD trees, which governs the discovery of the partitions. It also overviews the specifics of the algorithm for RD trees along with some practical guidance on bandwidth and order of polynomial selection. Section 4 shows the Monte Carlo simulation results with sharp regression discontinuity design for linear and nonlinear in running variable cases. Section 5 demonstrates the usefulness of the algorithm on datasets, collected by Pop-Eleches and Urquiola (2013). Section 6 extends the method to fuzzy RD designs. Section 7 concludes.

2 Regression Discontinuity Tree

With classical regression discontinuity design, researchers are interested in the causal effect of a binary treatment. Let $Y(1)$ denote the potential outcome, when a unit gets the treatment and $Y(0)$ if no treatment takes place. The observed outcome corresponding to the actual treatment status can be written as

$$Y = Y(D) = \begin{cases} Y(0), & \text{if} \quad D = 0, \\ Y(1), & \text{if} \quad D = 1\,. \end{cases}$$

[9]This is the average score within schools for incoming students. The transition score is calculated based on students' performance on the national test(s) and by their previous grades during classes 5-8.

Treatment assignment in sharp RD[10] is a deterministic function of a scalar variable, called the *running variable*, which is denoted by X. This paper considers the standard case, in which the treatment D is determined solely by whether the value of the running variable is above or below a *fixed* and *known* threshold c :

$$D = \mathbb{1}_c(x) = \mathbb{1}_{[c,\infty)}(x) \begin{cases} 1\,, & \text{if} \quad x \geq c \\ 0\,, & \text{otherwise} \end{cases}$$

Treatment heterogeneity comes in the form of additional characteristics. Let Z be a set of K random variables referring to the possible sources of heterogeneity. Z are pre-treatment variables, therefore they must not have any effect on the value of the running variable. Following the machine learning terminology, call these variables *features*.
This paper proposes a method to estimate, or in some cases approximate, the conditional average treatment effect function given by

$$\tau(z) = \mathbb{E}\left[Y(1) - Y(0) | X = c, Z = z\right] \tag{1}$$

This function can be continuous, discrete, or a mixture in Z. The proposed regression tree algorithm does not allow for such flexibility in each case but gives a step-function approximation when this CATE function is continuous in z. I will now introduce the basics of regression trees.

2.1 CATE in regression discontinuity tree

Regression trees – sometimes referred to as a partitioning scheme – allows one to construct a simple, intuitive, and easy-to-interpret step-function approximation to the CATE. A tree Π corresponds to a partitioning of the feature space. Partitioning is carried out by recursive binary splitting: 1) Split the sample into two sub-samples along one feature with a split value. If a unit has a larger value for the selected feature than the split value, then it goes to the first sub-sample, otherwise to the second sub-sample. 2) If needed, one repeats the split, but now one considers the already split sub-samples for the next split. This way the feature space is partitioned into mutually exclusive rectangular regions. These final regions are called '*leaves*' or '*partitions*', denoted by ℓ_j. A regression tree, Π has $\#\Pi$ leaves, $j = 1, \ldots, \#\Pi$, whose union gives back the complete feature space $\mathbb{Z}$.

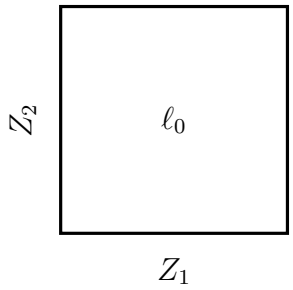
$$\Pi = \{\ell_1, \ldots, \ell_j, \ldots, \ell_{\#(\Pi)}\}, \quad \text{with} \quad \bigcup_{j=1}^{\#\Pi} \ell_j = \mathbb{Z}$$

[10]For fuzzy design, see Section 6

For illustrative purposes, consider only two features Z_1 and Z_2. Figure 1 shows three different trees with two representations.

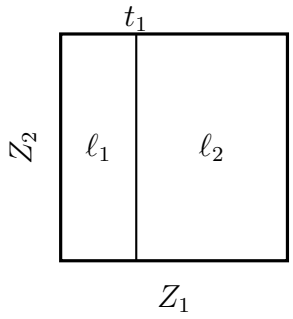
(1) **(2)**

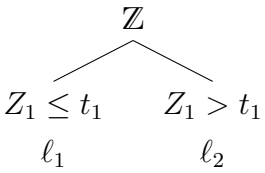
Partitioning **Tree structure**



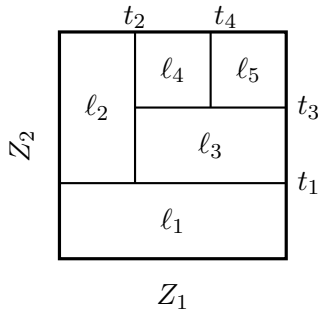
$\ell_0 = \mathbb{Z}$

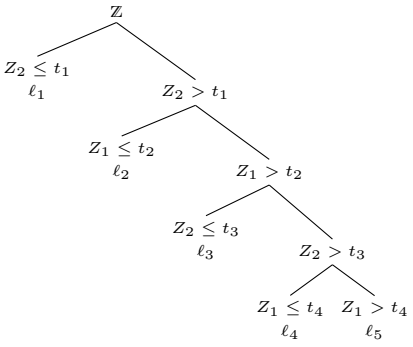
a) Π_0: homogeneous treatment effect, no partitioning





b) Π_1: two leaves, two treatment effects





c) Π_2: five leaves, five treatment effects

Figure 1: Different trees and their conditional average treatment effects

Column (1) shows the partitioning scheme: how the different partitions (or leaves) are split along the two features. Column (2) shows the tree structure: an intuitive interpretation using yes or no decisions, depending on the feature values and on the splitting values. Figure 1a) shows a tree, where there is only one leaf ℓ_0 containing all the units. This tree corresponds to a homogeneous treatment effect: no matter which values Z_1 or Z_2 takes, the treatment effect is always the same. In this case, the conditional average treatment effect is the same as the simple average treatment effect. Figure 1b) has two leaves: ℓ_1 and ℓ_2 resulting in two different treatment effects. Leaf ℓ_1 contains values with $Z_1 \leq t_1$ and ℓ_2 contains $Z_1 > t_1$, where t_1 is the splitting value. Note that Z_2 does not affect the partitioning and irrelevant with respect to treatment heterogeneity. Finally Figure 1c) shows a tree with five different leaves, resulting in five different treatment effects depending on both Z_1 and Z_2. In this case if one wants to find the treatment effect for a unit with $Z_1 = z_1$ and $Z_2 = z_2$, one needs to go through the decisions given by the tree. *Example*: $z_1 > t_3$ and $t_2 < z_2 \leq t_4$, corresponds to leaf ℓ_4. Note that the splitting values must satisfy $t_3 > t_1$, $t_1, t_3 \in Supp(Z_1)$ and $t_2, t_4 \in Supp(Z_2)$.

Recursive splitting provides rectangular regions for the different treatment effects, but never a continuous function. In the case of a continuous CATE, a simple tree offers only a step-function approximation. However, the tree structure ensures an intuitive decision-based interpretation of the treatment effects. Until Section 3, let us assume that the (true) tree Π is given. Using this known tree, the average treatment effect for leaf ℓ_j is defined as

$$\tau_j = \mathbb{E}\left[Y(1) - Y(0) | X = c, Z \in \ell_j(\Pi)\right] \tag{2}$$

To state the regression discontinuity tree approximation to the whole CATE function, let me introduce the indicator function for leaf ℓ_j.

$$\mathbb{1}_{\ell_j}(z;\Pi) = \begin{cases} 1\,, & \text{if} \quad z \in \ell_j(\Pi) \\ 0\,, & \text{otherwise} \end{cases}$$

The approximated conditional average treatment effect function provided by the regression discontinuity tree is given by

$$\tau(z;\Pi) = \sum_{j=1}^{\#\Pi} \tau_j \mathbb{1}_{\ell_j}(z;\Pi) \tag{3}$$

This CATE function – which incorporates the tree structure – links the treatment effects for each leaf. As the leaves represent rectangular partitions, this function is a step-function

approximation to the continuous CATE function. By the law of iterated expectation, this approximation has the property of $\mathbb{E}\left[\tau(Z) \mid \mathbb{1}_{\{c\}}(X), \mathbb{1}_{\ell_1}(Z), \ldots, \mathbb{1}_{\ell_{\#\Pi}}(Z)\right] = \tau(Z;\Pi)$. This means that at the threshold value ($X = c$) with the given tree structure, the expected value of the continuous CATE function over the leaves, is equal to the step-approximated CATE.

2.2 Identification of CATE in the sharp RD

To identify the conditional average treatment effect function for trees in sharp RD, the following assumptions are needed:

Identification assumptions

i) $\mathbb{E}[Y(1)|X = x, Z \in \ell_j(\Pi)]$ and $\mathbb{E}[Y(0)|X = x, Z \in \ell_j(\Pi)]$, exists and continuous at $x = c$ for all leaves in the tree.

ii) Let $f_j(x)$ denote the conditional density of x in leaf j. In each leaf j, c is an interior point of the support of $f_j(x)$.

Assumption i) states that the expected value of the potential outcomes conditional on the running variable in each leaf exists and continuous. It is required to identify the average treatment effects for all leaves. This assumption is similar to the classical RD assumption (see e.g., Imbens and Lemieux (2008)), but somewhat stronger, due to extension to the tree.[11] Assumption ii) ensures that the density for the running variable is well behaved: it has a positive probability below or above the threshold value within each leaf. This excludes cases when there are no values of the running variable on both sides of the threshold in a given leaf. Finally, in the RD literature, it is common to require the continuity of the conditional distribution functions – in this case, it extends to $f_j(x)$ to be continuous in x [12] – which is an implication of *"no precise control over the running variable"* (see e.g., Lee and Lemieux (2010)). In case, when local randomization around the threshold holds, the algorithm does not need this assumption. [13]

[11] But less restrictive if one assumes continuity in $Z = z$ as in e.g., Hsu and Shen (2019).

[12] One need to use the Bayes' Rule to show this, along with assumption i)

[13] *Note:* Although the used conditional average treatment effect function here is a step-function approximation, it can be a building block of a causal forest for sharp RD, which produces continuous condition average treatment effect. In this case, one needs further modification on the assumption for the conditional expectation and densities. A causal forest for RD is out of the scope of this current paper.

If these assumptions hold, the (step-function approximated) conditional average treatment effect given by a regression discontinuity tree is identified as

$$\begin{aligned}
\tau(z;\Pi) &= \sum_{j=1}^{\#\Pi} \tau_j \mathbb{1}_{\ell_j}(z;\Pi) \\
&= \sum_{j=1}^{\#\Pi} \left\{ \mathbb{E}\left[Y(1)|X=c, Z\in\ell_j(\Pi)\right] - \mathbb{E}\left[Y(0)|X=c, Z\in\ell_j(\Pi)\right] \right\} \; \mathbb{1}_{\ell_j}(z;\Pi) \\
&= \sum_{j=1}^{\#\Pi} \left\{ \lim_{x\downarrow c} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] - \lim_{x\uparrow c} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] \right\} \; \mathbb{1}_{\ell_j}(z;\Pi) \\
&= \mu_+(c,z;\Pi) - \mu_-(c,z;\Pi)
\end{aligned} \tag{4}$$

where

$$\begin{aligned}
\mu_+(x,z;\Pi) &= \sum_{j=1}^{\#\Pi} \mathbb{E}\left[Y(1)|X=x, Z\in\ell_j(\Pi)\right] \; \mathbb{1}_{\ell_j}(z;\Pi) \\
\mu_-(x,z;\Pi) &= \sum_{j=1}^{\#\Pi} \mathbb{E}\left[Y(0)|X=x, Z\in\ell_j(\Pi)\right] \; \mathbb{1}_{\ell_j}(z;\Pi)
\end{aligned} \tag{5}$$

refers to the conditional expectation function for (μ_+) above the threshold (treated) and (μ_-) below the threshold (untreated) units. That is, each τ_j is identified within its leaf in the usual way.

2.3 Interpretation of the estimand conditional on (un)observables

The conditional average treatment effect estimand in regression discontinuity designs is not as straightforward as in experimental designs or observational studies with the unconfoundedness assumption. To interpret the estimand, let me formalize the individual treatment effect as in Lee and Lemieux (2010),

$$Y(1) = Y(0) + \tau(Z,U)$$

where Z are the known observed covariates and U is unobserved heterogeneity in the individual treatment effect. In classical sharp RD setup[14], $\tau(Z,U)$ does not depend directly on the running variable X. Here, I will consider this simple case. Note that X, Z, and U can be correlated in this setup, thus individuals with characteristics of Z and U can have typical

[14]Simple assignment rule: $X \geq c$, the individual gets the treatment, otherwise not.

X values, but X does not directly influence the magnitude of the treatment effect. Naturally, individual treatment effects can not be observed as one can not assign the same unit to be treated and non-treated at the same time. Instead, one can identify a type of conditional average treatment effect, where Z and X are fixed and U is averaged out:

$$\tau(z) = \mathbb{E}\left[\tau(Z,U) \mid X = c, Z = z\right] = \mathbb{E}\left[Y(1) - Y(0) \mid X = c, Z = z\right] .$$

For purposes of interpretation, I consider the case when Z and U are discrete, but a similar argument applies to the continuous case. First, focus on the general case when no tree structure is used. CATE is identified through

$$\tau(z) = \lim_{x \downarrow c} \mathbb{E}\left[Y \mid X = x, Z = z\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y \mid X = x, Z = z\right] .$$

For the identifying equality to hold, the following extension of standard continuity conditions must hold:

i) $\mathbb{E}\left[Y(1)|X = x, Z = z\right]$ and $\mathbb{E}\left[Y(0)|X = x, Z = z\right]$, exists and continuous at $x = c$.

ii) Let $f(x \mid Z = z)$ denote the conditional density of x given $Z = z$. For each value of $z \in \text{Supp}(Z)$, c is an interior point of the support $f(x \mid Z = z)$.

Under these conditions, the CATE function is equal to,

$$\begin{aligned}
\mathbb{E}\left[Y(1) - Y(0) \mid X = c, Z = z\right] &= \mathbb{E}\left[\tau(Z,U) \mid X = c, Z = z\right] \\
&= \sum_u \tau(z,u) \mathbb{P}\left[U = u \mid X = c, Z = z\right] \\
&= \sum_u \tau(z,u) \frac{f(c \mid U = u, Z = z)}{f(c \mid Z = z)} \mathbb{P}\left[U = u \mid Z = z\right]
\end{aligned}$$

where $\mathbb{P}\left[\cdot \mid \cdot\right]$ denotes conditional probability and $f(\cdot \mid \cdot)$ denotes conditional density function. This formula is the exact analog of equation (5) in Lee and Lemieux (2010).

Thus, the CATE function is a particular kind of average treatment effect across individuals with covariate values $Z = z$. If the term $f(c|U = u, Z = z)/f(c \mid Z = z)$ were equal to 1, it would be the treatment effect for individuals with observed $Z = z$ averaged over the unobserved $U = u$ values. This is the case if the unobserved heterogeneity U is independent of the running variable X conditional on the covariates Z. More generally, the presence of the ratio $f(c|U = u, Z = z)/f(c \mid Z = z)$ implies the regression discontinuity estimand is instead a weighted average treatment effect. Within the subgroup $Z = z$, the weight is larger for individuals whose X value is ex-ante more likely to be close to the threshold

c based on their unobserved characteristics. The weights may be relatively similar across individuals, in which case the individual treatment effects would be closer to the CATE but, if the weights are highly varied and also related to the magnitude of the treatment effect, then the individualized treatment effects would be very different from the CATE. However, the weights across individuals are ultimately unknown, since we do not observe U. Thus, it is not possible to know how close the individualized treatment effects are to the CATE and it remains the case that the treatment effect estimated using an RD design is averaged over a larger population than one would have anticipated from a purely "cutoff" interpretation.

Finally, let me discuss the impact of using a regression tree representation in the interpretation of the CATE function. Following equation (2) from the paper, the leaf-by-leaf treatment effect can be similarly decomposed as

$$\begin{aligned}
\tau_j &= \mathbb{E}\left[Y(1) - Y(0) \mid X = c, Z \in \ell_j\right] \\
&= \mathbb{E}\left[\tau(Z, U) \mid X = c, Z \in \ell_j\right] \\
&= \sum_{z,u} \tau(z,u) \mathbb{P}\left[Z = z, U = z \mid X = c, Z \in \ell_j\right] \\
&= \sum_{z \in \ell_j, u} \tau(z,u) \frac{f(c \mid Z = z, U = u, Z \in \ell_j)}{f(c \mid Z \in \ell_j)} \mathbb{P}\left[Z = z, U = u \mid Z \in \ell_j\right] .
\end{aligned}$$

The interpretation remains similar, but with tree structure one needs to average over not only the unobserved characteristics ($U = u$), but over the observed characteristics within each leaf j as well.

Remarks:

i) If there is no unobserved heterogeneity in the treatment effect ($\tau(Z, U) = \tau(Z)$) then in the continuous case one can estimate the individualized treatment effects. With tree structure, weights are still present as the conditional densities are not necessarily same within leaf j for each values of z.

ii) In case the tree specification is correct in the sense that $\mathbb{E}\left[\tau(Z, U) \mid X = c, Z = z\right] = \mathbb{E}\left[\tau(Z, U) \mid X = c, Z \in \ell_j\right]$, then the interpretation is the same as if $\tau(Z, U)$ would be continuous in Z.

iii) If the tree is correctly specified and there is no unobserved heterogeneity in the treatment effect then the CATE via tree structure is the same as the individualized treatment effect.

2.4 Parametrization and estimation

The paper assumes q-th order polynomial functional form in X for each leaf to identify τ_j. Each conditional expectation function – $\mathbb{E}\left[Y(d)|X=x, Z\in\ell_j(\Pi)\right], d\in\{0,1\}$ – is given by a q-th order polynomial, which ensures a flexible functional form. [15] To formalize the parametrization of the conditional expectation function given by equation (5) first adjust X by c, and let $\boldsymbol{X}$ be the $(q+1)\times 1$ vector

$$\boldsymbol{X} = \left[1, (X-c), (X-c)^2, \ldots, (X-c)^q\right]'$$

For a given Π, one can then write

$$\mu_+(x,z;\Pi) = \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}_j^+\,, \qquad \mu_-(x,z;\Pi) = \boldsymbol{X}'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}_j^-$$

where, $\boldsymbol{\delta}_j^+ = \left[\alpha_j^+, \beta_1^+, \beta_2^+, \ldots, \beta_q^+\right]'$ and $\boldsymbol{\delta}_j^- = \left[\alpha_j^-, \beta_1^-, \beta_2^-, \ldots, \beta_q^-\right]'$ are a $(q+1)\times 1$ parameter vectors[16] and depends on the partitioning. Note that this definition allows for each leaf (thus group) to have different functional forms in X.

To estimate $\tau(z;\Pi)$ consider a sample $\mathcal{S}$, consisting of independent and identically distributed observations $(Y_i, X_i, Z_i)\,; i=1,\ldots,N$. The paper employs leaf-by-leaf estimation for the parameter vectors $\boldsymbol{\delta}_j^+$ and $\boldsymbol{\delta}_j^-$, using least squares.[17] The estimator for the parameters are given by

$$\hat{\boldsymbol{\delta}}_j^+ = \arg\min_{\boldsymbol{\delta}_j^+} \sum_{i\in\mathcal{S}}\left\{\mathbb{1}_c(X_i)\mathbb{1}_{\ell_j}(Z_i;\Pi)\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\delta}_j^+\right)^2\right\}$$
$$\hat{\boldsymbol{\delta}}_j^- = \arg\min_{\boldsymbol{\delta}_j^-} \sum_{i\in\mathcal{S}}\left\{\left[1-\mathbb{1}_c(X_i)\right]\mathbb{1}_{\ell_j}(Z_i;\Pi)\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\delta}_j^-\right)^2\right\}\,, \qquad \forall j$$

Using these parameter vectors and the identification equation for CATE (equation 4), the least squares estimator for conditional average treatment effect for regression discontinuity

[15] Nonparametric estimations such as local polynomial regression are not considered in this paper – mainly because of optimal criterion for growing a tree is more cumbersome in the presence of potentially multiple bandwidths – however in the case of strong non-linearity in X, I recommend using a restricted sample using a bandwidth (e.g., proposed by Imbens and Kalyanaraman (2012)), which is estimated on the whole sample.

[16] For RD the main parameter of interest is $\alpha_j^{\pm}$. $\beta^{\pm}$ should also be $\beta_1, j^{\pm}$, but I neglect j subscript for convenience.

[17] This method has the advantage of relative fast estimation. Computationally it is much more compelling than the two other alternatives: 1) joint estimation of the whole tree and 2) also include in one regression the treated and non-treated units. Although, these methods use milder assumptions during the search for the proper tree, but when estimating with these setups there is a need for inverting large sparse matrices (interactions of $\mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{X}'$), which can lead to computationally expensive methods and non-precise estimates.

tree is given by,

$$\hat{\tau}(z;\Pi,\mathcal{S}) = \hat{\mu}_{+}(c,z;\Pi,\mathcal{S}) - \hat{\mu}_{-}(c,z;\Pi,\mathcal{S}) = \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\,(\hat{\alpha}_{+,j} - \hat{\alpha}_{-,j})$$

Remark: The sample $\mathcal{S}$ is highlighted in this notation, due to later purposes to differentiate between estimates using different samples. Subscript i always refers to observations from sample $\mathcal{S}$, j index represents leaf j from tree Π and subscripts $+/-$ stands for above or below the threshold.

3 Discovering regression discontinuity tree

In this section, the assumption of a known tree is gradually relaxed. I approach this problem in three steps. Firstly, I introduce different distinct samples which are necessary to obtain an unbiased estimator of the CATE function, when using the regression tree algorithm. Here, I sketch some properties of the algorithm, which is detailed in the last step. Secondly, I analyze the criterion, which compares different trees. At this stage, I assume that these different trees are exogenously given. Finally, I show how the optimal tree is found by the regression tree algorithm, using the different samples and the proposed criterion.

3.1 Distinction of samples

An inherent problem of using only one sample for finding relevant sub-groups and estimating treatment effects is that it results in incorrect inference if there is no adjustment for multiple testing. (see, e.g., Romano and Shaikh, 2010 or Anderson, 2008)

Although regression tree algorithm controls for over-fitting in some way – as I will discuss in Section 3.3 – the estimate is biased in finite samples and disappears only slowly as the sample size grows. Athey and Imbens (2016) proposes '*honest regression tree*' approach to eliminate the bias from the estimated conditional average treatment effects in experimental settings or observational studies with the unconfoundedness assumption. By their definition, a regression tree is called 'honest' if it does not use the same information for growing the candidate trees as for estimating the parameters of that tree. This requires using two *independent* samples. The '*test sample*' ($\mathcal{S}^{te}$) is used for evaluating the candidate trees and the '*estimation sample*' ($\mathcal{S}^{est}$) for estimating the treatment effects. These samples are also used to derive and analyze the honest criterion for the regression discontinuity tree. In Section 3.3 I elaborate further on how the samples are used when growing a tree. Honesty has the implication that the asymptotic properties of treatment effect estimates within the partitions are the same as if the partition had been exogenously given, thus biases are eliminated and one can conduct inference in the usual way. The cost of the honest approach is the loss

in precision – less observation used – due to sample splitting (Athey and Imbens, 2016, p. 7353-7354).[18]

3.2 Criterion for RD tree

A natural – but in-feasible criterion – for evaluating the regression discontinuity tree would be minimizing the mean squared error of the estimated CATE on the test sample. Let a partition (Π) be exogenously given. The CATE function ($\hat{\tau}(Z_i; \Pi, \mathcal{S}^{est})$) is estimated on $\mathcal{S}^{est}$ and evaluated on $\mathcal{S}^{te}$. The in-feasible MSE criterion is

$$MSE_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = \frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \left\{ \left[\tau(Z_i) - \hat{\tau}(Z_i; \Pi, \mathcal{S}^{est}) \right]^2 - \tau^2(Z_i) \right\} \tag{6}$$

where N^{te} is the number of observations in the test sample. Note, in this formulation, there is an extra adjustment term, $\tau^2(Z_i)$ – a scalar, independent of Π. Thus, it does not have any effect on the results, but it facilitates theoretical derivations. Furthermore, let me emphasize that this in-feasible criterion utilizes both the estimation sample and the test sample in a way that observations are needed to be known for both samples.

Calculating this criterion for different exogenously given trees would allow one to find the tree, whose deviation from the true CATE function is the smallest in the test sample. The problem is $\tau(\cdot)$ is unknown, thus this criterion is in-feasible. Instead – following Athey and Imbens (2016) – I minimize the *expected* MSE over the test and estimation samples. This formulation has two advantages: i) it gives the best fitting tree for the *expected* test and estimation sample. This is favorable because when the tree is grown, both of these samples are locked away from the algorithm (see Section 3.3). ii) using this formulation, an estimable criterion can be derived for comparing trees in practice. The expected MSE criterion is given by

$$EMSE_\tau(\Pi) = \mathbb{E}_{\mathcal{S}^{te}, \mathcal{S}^{est}} \left[MSE_\tau(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) \right] \tag{7}$$

This paper advocates trees (Π), which gives the smallest $EMSE_\tau$ value from all the candidate trees. Based on Athey and Imbens (2016), this EMSE criterion can be decomposed into two

[18] With the honest approach one does not need to place any external restrictions on how the tree is constructed. In the literature, there are other papers, which use additional assumptions to get valid inference, which is also possible - but in my opinion a more restrictive approach. An example is Imai et al. (2013), which uses 'sparsity' condition: only a few features affect the outcomes.

terms,[19] which helps to evaluate why this criterion offers a good choice for selecting a tree.

$$EMSE_\tau(\Pi) = \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right] \tag{8}$$

This formulation highlights the trade-off between finding new different treatment effects – hence larger trees – and minimizing the variance of the estimated treatment effects. The expected value[20] of the squared CATE ($\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right]$) prefers trees which are larger, as the expected squared treatment effects grows as there are more leaves (or groups). On the other hand, any estimator for this term is increasing in the number of splits, which leads to select trees, that are too large, i.e. where the treatment effects are in fact the same in different leaves. This is called over-fitting the true tree. The first term, the expected value of the treatment effect variances, explicitly incorporates the fact that finer partitions generate greater variance in leaf estimates in finite samples. Therefore it prefers smaller trees, where the average variance of the estimated treatment effects is lower. Through this channel, this term offsets the over-fitting caused by the expected value of the squared treatments. Note that, the expected variance term may select larger trees if leaves (or groups) have the same treatment effect, but have lower expected variances.

A technical contribution of this paper is to provide estimators for the expected treatment variances and the expected squared treatment effects in the regression discontinuity setup. Here I only present the results, refer to Appendix B for the derivations.[21]
In order to analyze the proposed estimator, let me first introduce the following expressions. Write the model as

$$Y_i = \mathbb{1}_c(X_i)\mu_+(X_i,Z_i;\Pi) + (1-\mathbb{1}_c(X_i))\mu_-(X_i,Z_i;\Pi) + \epsilon_i$$

where ϵ_i is the idiosyncratic disturbance term. Furthermore, let

$$\hat{\sigma}^2_{+,j} = \frac{1}{N^{te}_{+,j} - q - 1}\sum_{i\in\mathcal{S}^{te}}\left[\mathbb{1}_c(X_i)\mathbb{1}_{\ell_j}(Z_i;\Pi)\hat{\epsilon}_i\right]^2,$$
$$\hat{\sigma}^2_{-,j} = \frac{1}{N^{te}_{-,j} - q - 1}\sum_{i\in\mathcal{S}^{te}}\left[\{1-\mathbb{1}_c(X_i)\}\mathbb{1}_{\ell_j}(Z_i;\Pi)\hat{\epsilon}_i\right]^2$$

[19] See the detailed derivations in Appendix A. To derive an estimable EMSE criterion, the assumption of $\mathcal{S}^{est}$ and $\mathcal{S}^{te}$ being independent of each other is key.

[20] Z_i refers to features from $\mathcal{S}^{te}$.

[21] For the derivations I have used two further simplifying assumptions: i) the share of observations within each leaf – the number of observations within the leaf compared to the number of observations in the sample – are the same for the estimation and test sample. ii) the shares of units below and above the threshold within each leaf are the same for the estimation and test sample. Asymptotically both assumptions are true.

be the within leaf variance estimators for the disturbance terms in leaf j with $N^{te}_{+,j}, N^{te}_{-,j}$ number of observations within the same leaf for above and below the threshold respectively. $\hat{\epsilon}_i$ are the OLS residuals. For simplicity, I assume the same finite variance within the leaves, when deriving these estimators.[22] (See Appendix D for extensions, which relax the finite variance assumption.) Furthermore, let the cross-product of the running variable above and below the threshold for leaf j be

$$M_{+,j} = \frac{1}{N^{te}_{+,j}} \sum_{i \in \mathcal{S}^{te}} \left(\boldsymbol{X}_i \boldsymbol{X}_i \mathbb{1}_{\ell_j}(Z_i; \Pi) \mathbb{1}_c(X_i) \right) ,$$
$$M_{-,j} = \frac{1}{N^{te}_{-,j}} \sum_{i \in \mathcal{S}^{te}} \left(\boldsymbol{X}_i \boldsymbol{X}_i \mathbb{1}_{\ell_j}(Z_i; \Pi)(1 - \mathbb{1}_c(X_i)) \right) .$$

Using these quantities, one can derive specifically scaled variance estimators for the parameter vectors in leaf j:

$$\mathbb{V}\left[\hat{\delta}^+_j\right] = \frac{\hat{\sigma}^2_{+,j} \hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} , \qquad \mathbb{V}\left[\hat{\delta}^-_j\right] = \frac{\hat{\sigma}^2_{-,j} \hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}$$

where $p^{est}_{+,j}$ and $p^{est}_{-,j}$ are the share of units above and below the threshold in the estimation sample within leaf j. (Specific scaling is explained in Remarks ii-iii), see below.)

Estimator for the expected variance of the treatment effects can be derived as an average of these variance estimators,

$$\hat{\mathbb{E}}_{Z_i} \left\{ \hat{\mathbb{V}}_{\mathcal{S}^{est}} \left[\hat{\tau}(z; \Pi, \mathcal{S}^{est}) \right] \Big|_{z=Z_i} \right\} = \frac{1}{N^{est}} \sum_{j=1}^{\#\Pi} \left\{ e_1' \left[\frac{\hat{\sigma}^2_{+,j} \hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j} \hat{M}^{-1}_{-,j}}{p^{est}_{-,j}} \right] e_1 \right\} \tag{9}$$

where $e_1 = [1, 0 \ldots, 0]$ is a $1 \times (q+1)$ selector-vector to choose the variances of the intercepts referring to the treatment effect.[23] N^{est} is the number of observations in the estimation sample, which is a result from the derivations (see Appendix B.1).

Remarks:

i) Although the variance of the treatment effects refers to the estimation sample, $\hat{\sigma}^2_{\pm,j}, M^{-1}_{\pm,j}, \forall j$ are calculated using only observations from the test sample. This is possible, as the estimation and the test samples are independent from each other, therefore the asymptotic estimators for these quantities are the same.

[22] Also called homoscedastic errors within each leaf – which refers to the variances of the errors *within* the leaves being the same. Note: that this is only assumed for within leaves and not for the whole partition, thus disturbance terms for all leaves (ϵ_i) do not need to be homoscedastic.

[23] In case of kink designs, the selector vector would choose the appropriate order of polynomial.

ii) To adjust the variance estimator in finite samples for the estimation sample, one only needs to use limited information from the estimation sample, namely the share of observations above and below the threshold $(p^{est}_{+,j}, p^{est}_{-,j})$.

iii) Using the leaf shares instead of the number of observations for above and below the threshold is possible, as the variance of the treatment effect estimators are the same for each observation within the leaf, therefore one can use summation over the leaves $(j = 1, \ldots, \#\Pi)$ instead of individual observations.

The estimator for the expected value of the squared true CATE (second part of equation 8), uses the squared of estimated CATE and corrects the resulting bias with the variance. The estimator uses only the test sample, apart from weights in the variance estimator.[24]

$$\hat{\mathbb{E}}_{Z_i}\left[\tau^2(Z_i;\Pi)\right] = \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{te}) - \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\} \quad (10)$$

The averaged squared treatment estimator prefers trees with many leaves. It is the sample analog for finding groups with different treatment effects. This term always increases as the number of leaves increases, while the average of the sum of squared treatment effects for two (or more) groups is always greater than the average of the sum of one averaged squared treatment effect. The second part is similar to the derived expected variance, but here the scaling for the average (N^{te}) comes from the test sample, as the estimator refers to the expected value over the test sample.[25] The weights of $p^{est}_{+,j}$ and $p^{est}_{-,j}$ comes from the estimation sample and they help the algorithm to avoid sample specific splits – see the discussion in Section 3.3.

Putting together the two estimators one gets the following estimable EMSE criterion for regression discontinuity trees:

$$\begin{aligned}\widehat{EMSE}_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi) = &-\frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{te}) \\ &+\left(\frac{1}{N^{te}}+\frac{1}{N^{est}}\right)\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\left(\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}\right)}{p^{est}_{+,j}} + \frac{\left(\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}\right)}{p^{est}_{-,j}}\right]e_1\right\}\end{aligned} \quad (11)$$

Minimizing this criterion leads to trees, where i) there is strong evidence for heterogeneity in

[24] See the derivations in Appendix B.2.

[25] An alternative estimator would be using the estimation sample only. However, my goal is to construct an EMSE estimator, which uses only the test sample's observation and only some additional information from the estimation sample, to ensure that during the tree building phase the estimation sample is locked away to get valid inference.

the treatment effects for different groups, and ii) penalize a partition that creates variance in leaf estimates. Furthermore, this criterion encourages partitions, where the variance of a treatment effect estimator is lower, even if the leaves have the same average treatment effect, thus finds features, which affect the mean outcome, but not the treatment effects themselves. Finally, let me compare the estimator for the EMSE criterion and the initial in-feasible MSE criterion. As the in-feasible MSE criterion uses the estimation sample to get an estimator for the CATE function and then evaluates it on the test sample, the estimator for EMSE criterion uses the observations from the test sample and only scales it with the number of observations (N^{est}) and share of units below and above the threshold for each leaf ($p^{est}_{\pm,j}$) from the estimation sample. This means there is only limited information needed from the estimation sample to calculate the EMSE criterion, but not individual observations. This property enables that the observation values from the estimation sample are locked away for the algorithm when searching for an optimal tree.

3.3 Finding EMSE optimal RD tree

Unit now, I have compared different, already given partitions using the proposed criterion. In this sub-section, I introduce the basic notations and steps to grow the EMSE optimal regression discontinuity tree, following the literature on classification and regression trees (CART) and honest causal regression trees. For more detailed description see, Breiman et al. (1984), Ripley (1996) or Hastie et al. (2011) on CART algorithms and Athey and Imbens (2015, 2016) on honest causal tree algorithm.

Finding the EMSE optimal honest RD tree has four distinct stages:

1. Split the sample into two independent parts.
2. Grow a large tree on the first sample.
3. Prune this large tree to control for over-fitting. This is carried out by cross-validation and it results in an EMSE optimal tree.
4. Use this EMSE optimal tree to estimate the CATE function on the independent estimation sample.

In the first stage '*honest*' approach randomly assigns the initial sample into two samples to achieve an unbiased CATE estimator. The first sample is called the *'training sample'* ($\mathcal{S}^{tr}$) and its observations are used to grow trees. The second, *'estimation sample'* has a special role. In general, it is locked away from the algorithm, but information on the number of observations is utilized during the tree building phase to control for finding training sample specific patterns. Observation values from the estimation sample are not used until the last stage. This division ensures valid inference for the CATE function in the fourth step. Figure 2 shows these two samples, which are used to grow a large tree and providing valid inference.

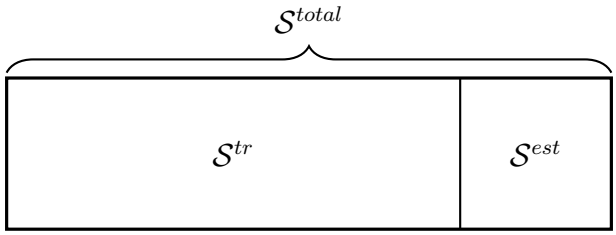


Figure 2: At stage 2, $\mathcal{S}^{tr}$ is used to grow large tree $\left(\hat{\Pi}^{large}\right)$

In the second stage, a large tree is grown using all the observations from the *training sample*. The algorithm recursively partitions the training sample along with the features. For each leaf the method evaluates all candidate features and their possible splits inducing alternative partitions, with the *'honest in-sample criterion'*: $\widehat{EMSE}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{est}, \Pi)$. This criterion uses additional information from the estimation sample. The treatment effects and the variances are estimated on the training sample only, but they are adjusted with the number of observations (N^{est}) and share of treated and non-treated units within each leaf from the estimation sample ($p^{est}_{\pm,j}$). N^{est} adjusts for the sample shares (how the initial sample is divided into two parts). This does not have a large impact on the in-sample criterion as the value is given by the first step and does not change during the partitioning. Using $p^{est}_{\pm,j}$ instead of $p^{tr}_{\pm,j}$ has larger implication in finite samples. It prevents the algorithm to choose such a feature and splitting value, which is only specific to the training sample. After the split is done, the algorithm iterates the procedure on the newly created leaves. The process repeats itself and stops if the in-sample-criterion does not decrease any further or the magnitude of the reduction is smaller than a pre-set parameter.[26] With this method, one gets a large tree $\left(\hat{\Pi}^{large}\right)$.

The resulting large tree is prone to over-fitting as $\widehat{EMSE}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{est}, \Pi)$ is not unbiased when one uses it repeatedly to evaluate splits on the training data. The bias comes from the fact that after the training sample has been divided once, the sample variance of observations in the training data within a given leaf is on average lower than the sample variance would be in a new, independent sample. This leads to finding features relevant, which are in fact irrelevant to the true CATE function. Thus using only $\widehat{EMSE}_\tau(\mathcal{S}^{tr}, \mathcal{S}^{est}, \Pi)$ is likely to overstate the goodness of fit as one grows deeper and deeper tree.

To solve for the over-fitting – in the third stage – cross-validation is used. The idea is to

[26]The algorithm accepts splits, where the in-sample-criterion decreases compare to a tree without the split. It is possible to specify a minimum amount of reduction in the in-sample-criterion, which by default is set to zero. Furthermore, the algorithm considers a split valid if the number of observations within the leaves is more than a pre-set value (typically 50 observations) for both the treated and control group. Finally, there are additional (optional) stopping rules implemented such as the maximum depth of the tree, the maximum number of leaves, the maximum number of nodes, or the maximum number of iteration.

split the training sample into two further parts: a sample where the tree is independently grown $\mathcal{S}^{(tr,tr)}$ and to a test sample $\mathcal{S}^{(tr,te)}$ where the EMSE criterion can be safely evaluated. This ensures that the tree grown on $\mathcal{S}^{(tr,tr)}$ is exogenous for $\mathcal{S}^{(tr,te)}$, thus the estimated EMSE criterion is unbiased. [27] Figure 3, shows the splitting of the original training sample into $\mathcal{S}^{(tr,tr)}$ and $\mathcal{S}^{(tr,te)}$.

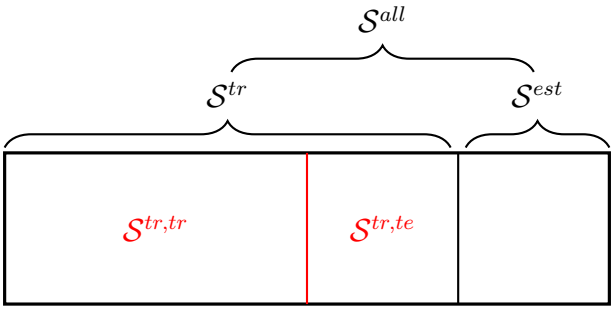


Figure 3: At stage 3, $\mathcal{S}^{tr,te}$ is used to evaluate the tree $\hat{\Pi}$ grown on $\mathcal{S}^{tr,tr}$

Note that, one needs to split the estimation sample as well for the accompanying information on the shares of treated and non-treated units to evaluate the EMSE criterion.
The EMSE optimal tree is found via cost-complexity pruning, which utilizes a complexity parameter (γ). The complexity parameter penalizes the number of leaves ($\#\Pi$) grown on the tree. The *'honest cross-validation criterion'* adds this penalty term to the original EMSE criterion,

$$\widehat{EMSE}_{cv}(\gamma) = \widehat{EMSE}_{\tau}(\mathcal{S}^{(tr,val)}, \mathcal{S}^{(est,val)}, \widehat{\Pi}) + \gamma \#\widehat{\Pi} \tag{12}$$

where, $\widehat{\Pi}$ is an estimator of the tree, grown on the samples of $\{\mathcal{S}^{(tr,tr)}, \mathcal{S}^{(est,tr)}\}$ and the EMSE criterion is evaluated on the independent sample pair of $\{\mathcal{S}^{(tr,te)}, \mathcal{S}^{(est,te)}\}$. To find the optimal complexity parameter (γ^*) – hence the EMSE optimal tree – one calculates the honest cross-validation criterion R times on the alternating test samples, which results in R criteria for each different candidate of γ.[28] Taking the average over the cross-validation

[27] The size of the samples are given by the number of folds (R) used in the cross-validation. $\mathcal{S}^{(tr,te)}$ has the smaller fraction: $N^{(tr,te)} = N^{tr}/R$, while the sample $\mathcal{S}^{(tr,tr)}$, which is used to grow the tree, contains the larger fraction of observations $N^{(tr,tr)} = (R-1)N^{tr}/R$. The estimation sample is split in the same way.

[28] The candidates of γ, coming from weakest-link pruning: using the large tree built on the whole training sample, γ values represent those penalty parameters which would result in a smaller tree for this large partition. During cross-validation, these scaled candidate γ values are used to prune back the trees. Scaling adjusts to the 'typical value' for the accompanied sub-tree. 'Long introduction for rpart package' gives an excellent overview on the technicalities of the cross-validation as well, which is available at `https://rdrr.io/cran/rpart/f/inst/doc/longintro.pdf` .

samples one can choose γ, which gives the minimum criterion value.[29]

$$\gamma^* = \arg\min_{\gamma} R^{-1} \sum_{cv=1}^{R} \widehat{EMSE}_{cv}(\gamma) \tag{13}$$

The final step of the third stage is to prune back the original large tree $\left(\hat{\Pi}^{large}\right)$ grown on the whole training sample with γ^* to get the optimal tree $\widehat{\Pi}^*$.
In the fourth stage, one uses the locked away estimation sample and the found tree structure $\widehat{\Pi}^*$ to estimate the CATE function for the regression discontinuity tree.

3.4 Refining honest tree algorithm for RD

In this subsection, I discuss the refinements of honest tree algorithms, which are needed for the estimation of the CATE function in regression discontinuity designs.
The main challenge for the RD algorithm takes place during the tree-building phase to find the optimal splitting values for each candidate feature. As the algorithm employs many regressions when considering each possible splits, the inversion of $M_{\pm,j}$ is computationally challenging. Instead of calculating the inverse each time, I use the Sherman-Morison formula to estimate $M_{\pm,j}^{-1}$. This iterative estimation enables the calculation of the inverse only once per splitting candidate feature.[30]

Another important detail of the honest algorithm is 'bucketing'. Following Athey and Imbens (2016), bucketing ensures that each candidate split has enough treated and non-treated units, thus there is no 'better' split value only due to adding treated or non-treated units, without the other. One should see bucketing as a smoother of the splitting criterion, as it groups the treated and non-treated units and prevents the splitting value to be a result of this unbalanced grouping of treated and non-treated units. I refine the classical causal tree algorithm[31], by carrying out the bucketing after the criterion is calculated and using the last valid split value instead of taking the average. This is an important nuisance as the criterion may vary too much without this modification for regression discontinuity trees.

Finally, there are two important issues specific to RD literature: selection of observations that are close to the threshold parameter to get 'local-randomization' and to get a precise conditional expectation estimate at the threshold from above and below. These issues, imply

[29] In the case of flat cross-validation criterion function it is well accepted to use 'one standard error rule': taking not the smallest value as the optimal, but the largest γ value which is within the one standard error range of the smallest value. This results in a smaller tree, which is easier to interpret and it filters out possible noise features, which would be relevant with the smallest cross-validation value.

[30] Note that there is a trade-off: if there are multiplicities in the value of a feature and it is not truly continuous, it may be faster to calculate the inverse for each candidate splitting value.

[31] Published at `https://github.com/susanathey/causalTree`

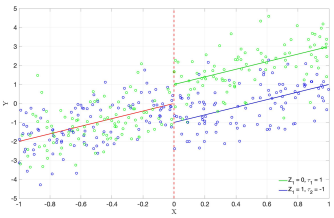
bandwidth selection procedure and choosing the order of polynomial during the estimation. Let us start with bandwidth selection. This paper does not offer a non-parametric method to estimate the conditional expectation function in the running variable, only parametric polynomials.[32] However, in practice a properly working solution is to use an under-smoothing bandwidth on the full sample, then restrict the used sample and employ the algorithm in this restricted sample. There is a recent discussion on selecting the order of polynomials used during the estimation (see, e.g., Gelman and Imbens (2019) or Pei et al. (2020)). [33] This paper offers a natural approach to select the order of polynomials: use the cross-validation procedure jointly with the complexity parameter to select q. As the estimated EMSE value is an unbiased estimator, it will lead to EMSE optimal order of polynomial selection as well.

4 Monte-Carlo simulations

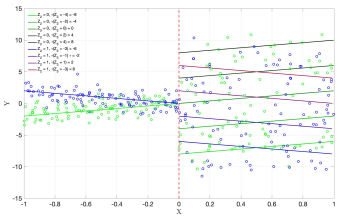
For Monte Carlo simulation, I created five different designs investigating different forms of heterogeneous treatment effects in RD. The first data generating process (DGP) is a simple example to demonstrate how the algorithm finds a simple tree structured DGP. Its simplicity comes from employing only two treatment effects which are defined by one dummy variable. The conditional expectation function (CEF) is linear and homogeneous across the leaves. DGP-2 imitates the step-function approximation nature of the algorithm: it has a continuous treatment effect function dependent on a single continuous variable, while the conditional expectation function is a linear function of another pre-treatment variable. DGP-3 to 5 revisit the simulation designs of Calonico et al. (2014) with non-linear conditional expectation function. I add heterogeneity to the treatment effects for DGP-3 and DGP-4, parallel to DGP-1 and DGP-2: two treatment effects defined by a dummy variable for DGP-3 and a continuous CATE for DGP-4. DGP-5 shows how the algorithm performs when there is no heterogeneity in the treatment effect. Figure 4 shows the different sharp RD designs.

[32] It would be an interesting research avenue to extend the EMSE criterion to non-parametric estimators as well. The algorithm could handle this extension naturally by including splitting the running variable as well, but the bias-variance trade-off would alter the behavior of the criterion.

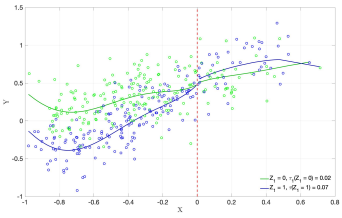
[33] The main recommendation of Gelman and Imbens (2019) is to use low order (local) polynomials to avoid noisy estimates. Pei et al. (2020) proposes a measure that incorporates the most frequently used non-parametric tools to select the order of polynomial.



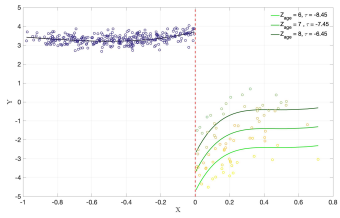
DGP-1: Two treatment effects with homogeneous CEFs



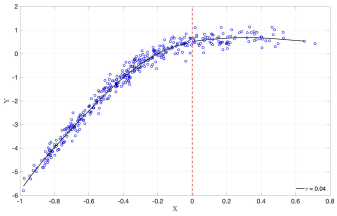
DGP-2: Heterogeneous CEF with continuous CATE



DGP-3: Two treatment effects with non-linear CEF



DGP-4: Continuous CATE with non-linear CEF



DGP-5: Homogeneous treatment effect with non-linear CEF

Figure 4: Monte Carlo simulation designs

During the simulations, I use three different sample sizes: $N = 1,000$; $5,000$ and $10,000$ to investigate the effect of the sample size on the algorithm. As the method splits the initial sample, I use half of the observations for the training and the other half for the estimation sample. I use $MC = 1,000$ Monte Carlo repetition and the variation comes from a normally distributed disturbance term, $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, whereas the features are uncorrelated[34]. For DGP 1 and 2, I use $\sigma_\epsilon^2 = 1$ and for DGP 3-5, $\sigma_\epsilon^2 = 0.05$.

Next, I discuss the main features of the proposed DGPs. For the complete specification refer to Appendix E.

DGP-1: Imitates a simple tree structure: there are two distinct treatment effects, conditioning on one binary variable. There is also an additional irrelevant binary variable. Both of them generated by using the probabilities of $P(Z_k = 1) = 0.5\,, k = \{1, 2\}$.

- $\tau(Z_1 = 1) = 1$, $\tau(Z_1 = 0) = -1$, number of features: 2

DGP-2: The second design follows Athey and Imbens (2016), who uses heterogeneous conditional expectation function along with continuous treatment effect. DGP-2 is modified for sharp RD and uses four different features: two binary (Z_1, Z_2 with $P(Z_1 = 1) = P(Z_2 = 1) = 0.5$) and two continuous ($Z_3, Z_4 \sim U(-5, 5)$) variables.

[34]This means, during the simulations, there are no issues with (highly) correlated features, which would alter the stability of the resulting trees.

The conditional expectation is a function of Z_2 along with the running variable, but has no effect on the magnitude of the treatment. CATE is a linear function of Z_3. This design shows a clean behavior for the step-function approximation, while allowing heterogeneity in the conditional expectation function.

- $\tau(Z_3) = 2\,Z_3$, number of features: 4

The last three designs investigate the performance of the algorithm when the conditional expectation function is non-linear. These setups use the functional forms proposed by Calonico et al. (2014) and imitate different RD applications. This exercise exhibits how the performance of the algorithm alters compared to the linear cases. To compare the behavior of the method I induce heterogeneity in the treatment effects similarly as in DGP 1 and 2, but add more potential pre-treatment variables.

DGP-3: Imitates Lee (2008) vote-shares application. I assume two treatment effects with different conditional expectation functions for the leaves. I use 52 dummy variables representing political parties and different states. The artificial political party dummy (Z_1) is relevant and affects on both treatment and the functional form. Artificial state variables are irrelevant.

- $\tau(Z_1 = 1) = 0.02$, $\tau(Z_1 = 0) = 0.07$, number of features: 52

DGP-4: Follows Ludwig and Miller (2007), who studied the effect of Head Start funding to identify the program's effects on health and schooling. I assume a continuous treatment effect based on the age of participants (Z_1), while adding (irrelevant) dummies representing different continents.

- $\tau(Z_1) = -0.45 - Z_1$, number of features: 7

DGP-5: An alternative DGP by Calonico et al. (2014), which adds extra curvature to the functional form. This design is the same as in Calonico et al. (2014), thus there is only one homogeneous treatment effect.

- $\tau = 0.04$, number of features: 52

To evaluate the performance of the algorithm I am using three different measures. The first measure investigates, whether the proposed estimable EMSE criterion is a good proxy to minimize the ideal in-feasible criterion (equation 6). For transparent comparison, I calculate this in-feasible criterion on a third independent evaluation sample, containing $N^{eval} = 10,000$ observations. The criterion is calculated on this evaluation sample, and the CATE estimator comes from the tree, which is grown on the training sample and estimated on the estimation sample. The Monte Carlo average of this estimate is reported as "*inf. MSE*". The second

measure is the average number of leaves on the discovered tree ($\#\hat{\Pi}$). DGP-2 and 4 with continuous CATE function should have an increasing number of leaves as one increases the number of observations. This would imply proper step-function approximation nature of the algorithm as more observations allow the algorithm to split more along with the relevant feature. For DGP-1, 3, and 5 the number of leaves should be the same as the number of distinct treatment effects in the true DGP. This measure may be misleading in cases when the algorithm finds different treatment effects, but the conditioning variables are not the same as in the true DGP (e.g., in DGP-1 the algorithm splits with Z_2 instead of Z_1, which would result in the same number of leaves, but not finding the true DGP). Therefore I also calculate the percent how many times the true DGP is found, when the DGP has a tree structure. (For DGPs with continuous CATE this is measure is not reported, as the algorithm only provides a step-function approximation of the true CATE.). Table 1. reports the results on the algorithm performance.

DGP	N	inf. MSE	$\#\hat{\Pi}$	DGP found (%)
DGP-1	$N = 1,000$	0.0620	2.00	100%
	$N = 5,000$	0.0135	2.04	96%
	$N = 10,000$	0.0065	2.04	96%
DGP-2	$N = 1,000$	9.3103	2.00	-
	$N = 5,000$	1.3852	7.72	-
	$N = 10,000$	0.9233	11.68	-
DGP-3	$N = 1,000$	0.0013	1.00	0%
	$N = 5,000$	0.0003	2.00	100%
	$N = 10,000$	0.0001	2.00	100%
DGP-4	$N = 1,000$	1.3904	1.00	-
	$N = 5,000$	0.4160	3.00	-
	$N = 10,000$	0.2013	4.92	-
DGP-5	$N = 1,000$	0.0007	1.00	100%
	$N = 5,000$	0.0002	1.03	97%
	$N = 10,000$	0.0001	1.02	98%

Table 1: Monte Carlo averages for performance measures
Number of true leaves: $\#\Pi_{DGP-1} = 2$, $\#\Pi_{DGP-3} = 2$, $\#\Pi_{DGP-5} = 1$
Algorithm setup: using the smallest cross-validation value to select γ^*,
$q = 1$ for DGP 1 and 2 and $q = 5$ for DGP 3,4 and 5.

From Table 1 one can see that the algorithm works considerably well. The infeasible MSE is decreasing in N for each setup. This supports the theoretical claim that the estimable EMSE

criterion is a proper proxy for the infeasible MSE, thus the resulted tree is MSE optimal in this sense. The average number of leaves on the discovered trees reflects the expectations. For DGP-2 and 4, where the CATE is continuous the average number of leaves is increasing in N. Note that the algorithm performs better when the CEF is linear compared to the non-linear case. For DGP-1, 3, and 5 the average number of leaves reflects the true number of leaves for the DGPs with one exception: for DGP-3 with $N = 1,000$. In this case, the algorithm does not split but gives a homogeneous treatment effect instead of the two distinct treatment effects. The measure of DGP found (%) reflects that the algorithm does not split along irrelevant variables but along relevant variables. Finally, results in Table 1 show that the algorithm is rather conservative in discovering different treatment effects and a data intensive method. In the case of DGP-1 and DGP-5, the signal-to-noise ratio is relatively high and with $N = 1,000$ it does not discover any irrelevant features (only the true DGP) – however due to randomness it should in some cases. DGP-3 on the other hand has a relatively low signal-to-noise ratio, and with $N = 1,000$ it never discovers the true DGP. Increasing the number of observations solves this problem. $N = 5,000$ observations are enough for DGP-1 and 5, but it takes $N > 10,000$ for DGP-3, showing the data intensity of the method.

Another important result for the regression discontinuity tree is Monte Carlo evidence on providing valid inference. I calculate the average bias and the actual 95% confident interval (CI) coverage for each leaf. Table 2 reports the Monte Carlo average of the bias for each leaf $\left(MC^{-1}\sum_{mc=1}^{MC}(\tau_j - \hat{\tau}_{j,mc})\right)$ and the actual 95% CI coverage for the different leaves conditionally whether the algorithm found the true DGP. I report only DGPs, which has a tree structure, as in cases of continuous CATEs, the leaves are varying due to different splitting values, making the reporting and aggregation over the Monte Carlo sample non-trivial.[35]

[35]Note that for continuous CATE the treatment effect conditional on the leaf is still an unbiased estimator for the given feature partition and has proper standard errors, however simple aggregation by the Monte-Carlo simulation distorts these properties.

DGP 1	Leaf	$\ell_1 : \tau_1(Z_1 = 1) = 1$		$\ell_2 : \tau_1(Z_1 = 0) = -1$	
	Estimates	average bias	actual 95% CI coverage	average bias	actual 95% CI coverage
	$N = 1,000$	-0.0121	0.95	-0.0155	0.95
	$N = 5,000$	-0.0015	0.95	-0.0022	0.94
	$N = 10,000$	0.0009	0.96	0.0003	0.95
DGP 3	Leaf	$\ell_1 : \tau_1(Z_1 = 0) = 0.07$		$\ell_2 : \tau_1(Z_1 = 1) = 0.02$	
	Estimates	average bias	actual 95% CI coverage	average bias	actual 95% CI coverage
	$N = 1,000$	-	-	-	-
	$N = 5,000$	0.0002	0.94	0.0000	0.95
	$N = 10,000$	-0.0000	0.95	0.0004	0.96
DGP 5	Leaf	Homogeneous Treatment, $\tau = 0.04$			
	Estimates	avgerage bias		actual 95% CI coverage	
	$N = 1,000$	-0.0001		0.95	
	$N = 5,000$	0.0001		0.96	
	$N = 10,000$	0.0004		0.95	

Table 2: Estimated Monte Carlo average for bias and actual 95% confidence intervals coverage for each leaves for tree structured DGPs, conditional on DGP is found

Note: For DGP-3, with $N = 1,000$, there is no case when the true DGP is found, thus no values are reported.

Table 2 shows that the average bias is decreasing in N for each leaf individually (at least up to 3 digits), similarly to the infeasible MSE, which is averaged over these leaves. The actual 95 % CI coverage reflects properly the nominal value. These results provide evidence of valid inference for the estimated CATE function.

5 Heterogeneous effect of going to a better school

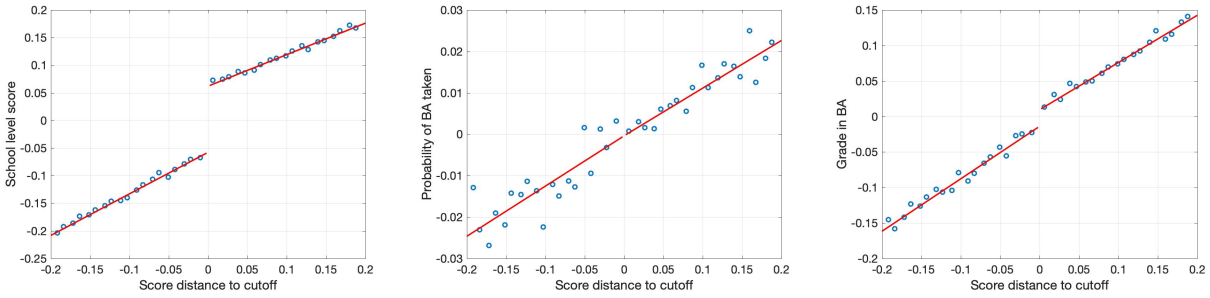
To show how the algorithm works in practice, I replicate and augment the heterogeneity analysis of Pop-Eleches and Urquiola (2013) on the effect of going to a better school. Furthermore, I relate my results to Hsu and Shen (2019).

In Romania, a typical elementary school student takes a nationwide test in the last year of school (8th grade) and applies to a list of high schools and tracks. The admission decision is entirely dependent on the student's transition score, an average of the student's performance on the nationwide test, grade point average, and order of preference for schools.[36] A student

[36]Grades on the nationwide test are from 1-10, where 5 is the passing score on each test. Grade point average is an average of the past years' course grades for different disciplines. Order of preference for schools

with a transition score above a school's cutoff is admitted to the most selective school for which he or she qualifies. Pop-Eleches and Urquiola (2013) use a large administrative dataset (more than 1.5 million observations) and a survey dataset (more than 10,000 observations) from Romania to study the impact of attending a more selective high school during the period of 2003-2007. Based on the administrative dataset, they find that attending a better school significantly improves a student's performance on the Baccalaureate exam,[37] but does not affect the exam take-up rate.

Figure 5 summarizes the classic mean RD results from Pop-Eleches and Urquiola (2013). In all three graphs, the horizontal axis represents the running variable, which is a student's standardized transition score subtracting the school admission cut-off. The vertical axis in Figure 5a) represents the *peer quality*, that each admitted student experiences, when going to school. Peer quality is defined as the average transition score for the admitted students in each school. This indicates that the higher the level of average transition score is (e.g., the admitted students performed great in the nationwide test), the better the peer quality. Figure 5b) shows the probability of a student taking the Baccalaureate exam, while Figure 5c) plots the Baccalaureate exam grade among exam-takers. In all outcomes, school fixed effects are used as in Pop-Eleches and Urquiola (2013), thus the vertical axis is centered around 0 for all plotted outcomes. Both left and right graphs show a jump in the average outcome at the discontinuity point, but the jump in the exam-taking rate is quite noisy and seemingly insignificant.



(a) Average transition score (b) Baccalaureate taken (c) Baccalaureate grade

Figure 5: Bin-scatter for main (pooled) RD results of Pop-Eleches and Urquiola (2013), using school fixed effects

is a list submitted by the student before the nationwide test, showing their preferences for the schools that they apply.

[37]Marks in BA Exam vary from 1-10, where there are multiple disciplines, wherein each, one needs to score above 5 and achieve a combined score of more than 6 to pass the BA Exam.

5.1 Revisiting heterogeneity analysis of Pop-Eleches and Urquiola (2013)

First, I revisit Pop-Eleches and Urquiola (2013) heterogeneity analysis on the intent-to-treat effects using peer quality (level of school average transition score) and the number of schools in town as the sources of heterogeneity, using the administrative data between 2003 to 2005.[38] Similarly, I restrict the sample to observations that lie within the ± 0.1 interval of the admission cutoff for the running variable and I use the same linear specification. Pop-Eleches and Urquiola (2013) inspect heterogeneity in the treatment effect with pre-specified sub-samples. The first two sub-samples are differentiated by the level of peer quality effect. Pop-Eleches and Urquiola (2013) investigate treatment effects for students in the top and bottom tercile for the school level average transition score. The second analysis focuses on the number of schools in town and creates groups defined by having i) four or more schools in towns, ii) three schools or iii) two schools only. Instead of using these pre-specified (ad-hoc) groups, I use the algorithm to identify the relevant groups and split values. I also use these two variables[39] to explore the heterogeneity but use them simultaneously allowing for finding different non-linear patterns in the treatment effect. See more details about these variables in Appendix G.

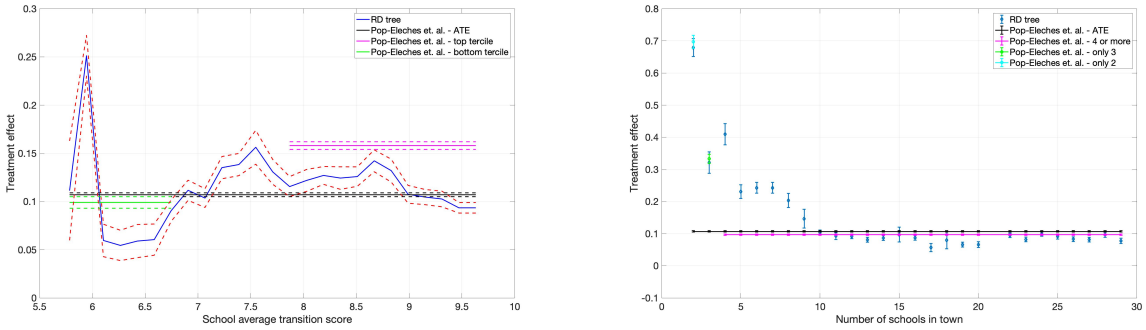
Let us consider the peer quality effect as the outcome, which is measured by the average transition score at their respective school. Pop-Eleches and Urquiola (2013) find significant positive treatment effects in all five groups. The regression discontinuity tree algorithm finds a much more detailed tree, containing 24 leaves, which is an indication of a continuous CATE function. Instead of showing a large tree, Figure 6 shows the marginalized treatment effects along the two variables.[40] Figure 6a) shows the treatment effects conditional on the level of school average transition score.[41] The blue line represents the CATE function found by the algorithm, the black line shows the overall average treatment effect, while the green and pink lines show the treatment effects reported by Pop-Eleches and Urquiola (2013) for the bottom and top tercile. Figure 6b) shows the heterogeneity in the treatment effects along with the number of schools. Similar to the previous plot, the different colored error bars show the treatment effects for the different models.

[38] Referring to Table 5 in Pop-Eleches and Urquiola (2013, p. 1310). See more details in Appendix G.

[39] I add dummy variables as well to search for a certain number of schools in the town.

[40] I have calculated the treatment effect for each observation then averaged them over the non-plotted variable. In the case of the number of schools, I take students with the same number of schools in town and average them along with the level of school average transition score.

[41] I used 50 equal-sized bins to group school average values.



(a) Level of school avg. transition score

(b) number of schools

Figure 6: CATE for peer quality, intent-to-treat effects, using school fixed effects, standard errors are clustered at student level

It is interesting to compare the algorithm's result (blue line) to Pop-Eleches and Urquiola (2013) results (green and pink lines). Figure 6a) shows that the 'bottom tercile' (green line) effect should be decomposed into two further parts: students with the lowest scores have high treatment effects, but students above score 6, but below 6.8 face the lowest treatment effects. This indicates a different mechanism of the treatment for these groups and aggregating them to the bottom tercile may lead to misleading suggestions. Conditioning the treatment effects on the number of schools results in the same conclusion for two and three schools,[42] but as Figure 6b) shows the treatment effect suggested by the algorithm is still higher than the average for towns with 4-9 schools and it is significantly lower for towns with 18-20 schools.

Investigating the treatment effect on the probability of taking the Baccalaureate exam – in contrast with Pop-Eleches and Urquiola (2013), who do not find significant treatment effects – the algorithm discovers a group where there is a significant negative effect on the exam-taking rate.

[42]The treatment effects are not the same as the algorithm uses only half the sample to estimate the CATE, thus the blue line is not varying exactly around the pink line (or the black/green lines).

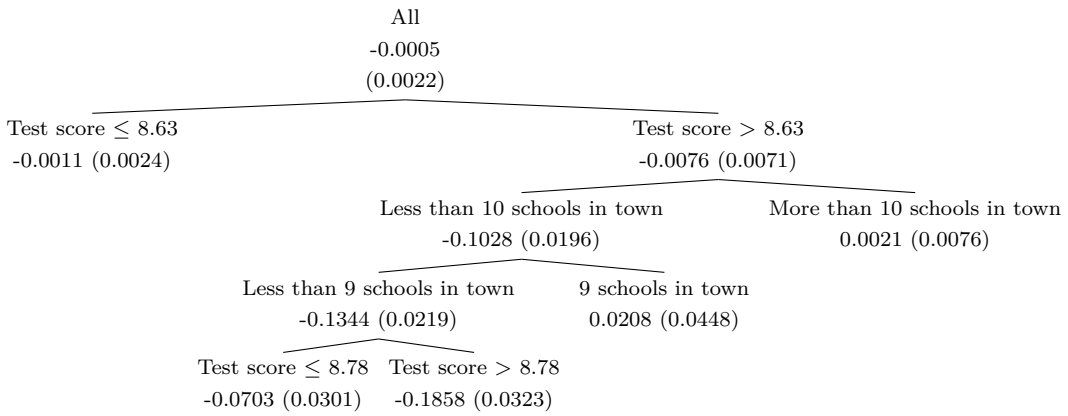


Figure 7: Conditional treatment effects for probability of taking Baccalaureate exam, intent-to-treat effects, using school fixed effects, standard errors are clustered at student level

Although the majority of the discovered groups have non-significant treatment effects, all these splits are needed to find the group which has a significant negative 19% treatment effect on the probability of taking the BA exam. As this value is surprisingly high, a researcher or policymaker may want to understand the background of this (sub-)population. The group is defined as students whose level of school average transition score is above 8.78 (top 10%) and in their town, there are less than 9 schools. Thus these students are admitted into an extremely competitive school, but there are few (or no) outside options to change school within the town. Overall, there are more than 20,000 cases that fall into this category. This result is aligned with the negative peer effect that Pop-Eleches and Urquiola (2013) report. Namely, on a distinct survey data set they find evidence that comparatively less talented students in competitive schools are less likely to go and take the Baccalaureate exam.

Last, the heterogeneity found by the algorithm in the value of Baccalaureate grade is the simplest as there are only two relevant groups. One group contains students whose school average transition score is above the median (to be exact, 7.4, which is the 44-th percentile in the sample). These students can expect a 0.0282 (0.0054) higher exam grade on the Baccalaureate exam if going to a better school, while students below this splitting value can expect only 0.0152 (0.0061) higher Baccalaureate exam grade. The algorithm does not split further, thus providing no further evidence on heterogeneity across the number of schools in town within these groups.[43]

Finally, let me relate these results to Hsu and Shen (2019). They search for heterogeneity using peer quality as the potential source and find strong evidence for the exam-taking rate (under 1% p-values) and weak evidence for BA grade (around 10% p-values) among schools.

[43]If one only uses the number of schools to find heterogeneity, the algorithm finds different treatment effects, but jointly it is non-relevant. See more details in Appendix G, Table 7.

Although they restrict their sample to towns with two or three schools and estimate the local average treatment effect, the conclusion is the same, values of school-level average test scores have an impact on the level of treatment effect.

5.2 Exploring heterogeneity in survey-based dataset

To explore treatment effect heterogeneity and show how the algorithm performs, when there are many covariates with potential non-linearities, I use the survey dataset from 2005-2007. This sample contains fewer observations, but a rich variety of socio-economic factors (e.g., gender, ethnicity, education, accessibility of internet or phone), school characteristics (e.g., novice teacher among teachers, highly certified teachers in schools), and study behavior-specific questions (e.g., parents pay for tutoring, parents help students, the student does homework every day, peer ranking, teacher characteristics). In the survey, there are only 135 schools located in 59 towns with 2 to 4 schools, and a questionnaire was administered between 2005 to 2007. Overall, I use 29 different features to search for heterogeneity. As the survey corresponds to later years, the data includes only observations on the level of school average transition scores, but not on the other two outcomes. See more detailed description in Appendix G.

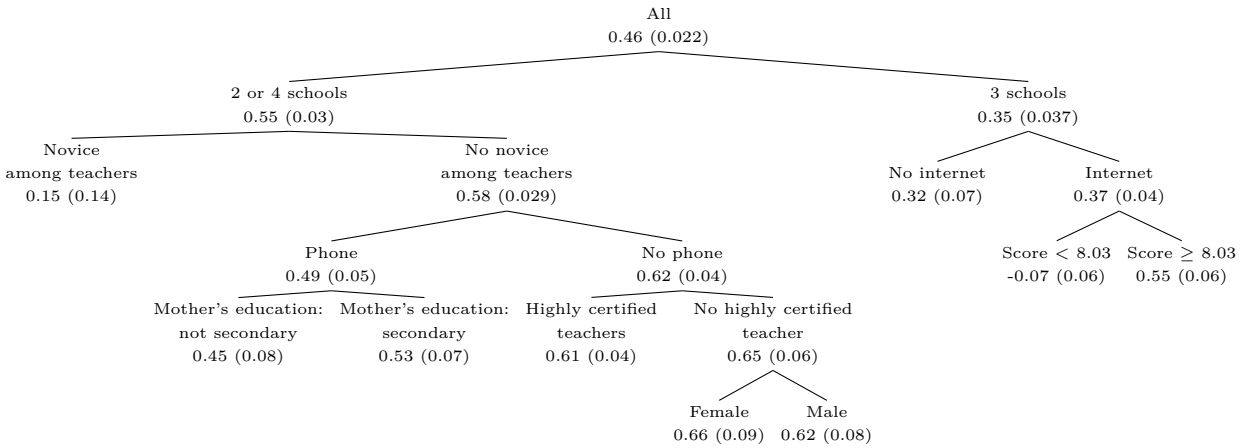


Figure 8: Exploring heterogeneous groups for peer quality - intent-to-treat effects, standard errors are in parenthesis and clustered on student level.

The fitted tree, shown in Figure 8, suggests an informative result: for towns with 2 or 4 schools, admitted students (on average) have 0.55 higher scores. If one goes further, the tree suggests that when having a novice among teachers (less than 2 years of experience), the treatment effect may disappear (although only 319 cases fall into this category). It is interesting that having a phone would somewhat reduce the peer quality effect in 2005-07 but it should also be noted that during the studied time-period it was not common for students

aged 12-14 to have phones. I also find interesting splits with respect to i) education of the mothers, ii) if there are teachers with the highest state certification in the school, and iii) if the student gender is male or female. These are indeed interesting splits but are statistically non-distinguishable.[44] Heterogeneity among groups in the other branch is more informative and more robust. If there are three schools and accessibility of the internet, the level of school average transition scores is an important split to identify a group. For schools, which have student scores above the bottom tercile (8.03 is the 35% percentile in this sample), the peer quality effect is similar to students with 2 or 4 schools in town (0.55 higher scores). However, for students in these schools with internet access at home, but below the bottom tercile, the peer quality effect is insignificant. This suggests potential segregation for this discovered group and encourages the researcher or policymaker to make further investigation on this specific group. Finally, let me note that the results are quite robust to randomization of the observations in the training/estimation sample. Some of the splits may vary but the main conclusion is similar in most of the cases.

6 Extension to fuzzy designs

The method can be extended to fuzzy designs as well, where the probability of treatment needs not to change from 0 to 1 at the threshold and can allow for a smaller jump in the probability of assignment.

Let me use a distinct variable T for getting the treatment in the case of fuzzy design. As the probability does not change from 0 to 1 at the threshold, there are different types of participants, depending on whether they are subject to the treatment or not. Compliers are units that get the treatment if they are above the threshold but do not get the treatment if they are below: $T(1) - T(0) = 1$. Always takers get the treatment regardless of whether they are below or above the threshold, while never takers never take the treatment regardless of the threshold value. For both behaviors, the following applies $T(1) - T(0) = 0$. As in classical fuzzy RD, I eliminate by assumption defiers, who do not take the treatment if above the threshold and take the treatment if below the threshold.

Fuzzy RD identifies treatment effects for compliers, thus extending the algorithm to fuzzy designs resulting in conditional local average treatment effects (CLATE). To identify CLATE, the following assumptions are needed:

Identifying assumptions of CATE in fuzzy RD

i) $\lim_{x \downarrow c} \mathbb{P}[T = 1|X = x] \geq \lim_{x \uparrow c} \mathbb{P}[T = 1|X = x]$

[44] As the cross-validation criterion is quite flat with the one standard error rule, these splits are pruned back.

ii) $\mathbb{E}\left[Y(d) \,|\, T(1) - T(0) = d', X = x, Z \in \ell_j(\Pi)\right]$ exists and continuous at $x = c$ for all pairs of $d, d' \in \{0, 1\}$ and for all leaves j in the tree.

iii) $\mathbb{P}\left[T(1) - T(0) = d \,|\, X = x, Z \in \ell_j(\Pi)\right]$ exists and continuous at $x = c$ for $d \in \{0, 1\}$, $\forall j$ and for all leaves j in the tree.

iv) Let, f_j denotes the conditional density of x in leaf j. In each leaf j, c must be an interior point of the support of $f_j(x)$.

Identification assumptions are similar to classical fuzzy RD, but it needs to be valid within each leaf. Assumption i) rules out defiers as it requires a non-negative discontinuity in the probability of taking the treatment around the threshold. This is not only an assumption, but a built-in restriction for the algorithm. If this condition's sample analogue is not satisfied, it is not considered as a valid split. Assumptions ii) and iii) ensure the existence and continuity of the expected potential outcomes at the threshold value for always-takers, compliers and never-takers with respect to the running variable within each leaf, while assumption iv) ensures that the conditional density of x for each leaf is well behaving, similarly to sharp RD.
Under these assumptions, the CLATE for RD tree is identified as

$$\begin{aligned}
\tau_{FRD}(z;\Pi) &= \frac{\lim_{x\downarrow c} \mu_+^y(x,z;\Pi) - \lim_{x\uparrow c} \mu_-^y(x,z;\Pi)}{\lim_{x\downarrow c} \mu_+^t(x,z;\Pi) - \lim_{x\uparrow c} \mu_-^t(x,z;\Pi)} \\
&= \frac{\mu_+^y(c,z;\Pi) - \mu_-^y(c,z;\Pi)}{\mu_+^t(c,z;\Pi) - \mu_-^t(c,z;\Pi)} \\
&= \frac{\tau^y(z;\Pi)}{\tau^t(z;\Pi)} \\
&= \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\ \frac{\alpha_{+,j}^y - \alpha_{-,j}^y}{\alpha_{+,j}^t - \alpha_{-,j}^t}
\end{aligned}$$

where, similarly to sharp RD, I use parametric functional forms for approximating the conditional expectation functions for both the participation and outcome equations below and above the threshold,

$$\mu_+^t(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}_j^{-,t} \ , \qquad \mu_+^y(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}_j^{-,y} \ ,$$

$$\mu_-^t(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}_j^{+,t} \ , \qquad \mu_-^y(x,z;\Pi) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\boldsymbol{\delta}_j^{+,y} \ ,$$

$$\boldsymbol{\delta}_{j,\pm}^t = \left[\alpha_{j,\pm}^t, \beta_{j,1,\pm}^t, \ldots, \beta_{j,p,\pm}^t\right]' \ , \qquad \boldsymbol{\delta}_{j,\pm}^y = \left[\alpha_{j,\pm}^y, \beta_{j,1,\pm}^y, \ldots, \beta_{j,p,\pm}^y\right]'$$

The sample estimates for fuzzy design are provided in Appendix C.

Using the same logic to find the optimal EMSE tree, I minimize the expected mean squared error function over the estimation and test sample. In the case of homoscedastic disturbance terms within each leaf, the estimable EMSE criterion for fuzzy designs is given by

$$\begin{aligned}\widehat{EMSE}_{FRD}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = &- \frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}^2_{FRD}(Z_i; \Pi, \mathcal{S}^{te}) \\ &+ \left(\frac{1}{N^{te}} + \frac{1}{N^{est}} \right) \sum_{j=1}^{\#\Pi} e_1' \left(\frac{\mathcal{V}_{+,j}}{p^{est}_{+,j}} + \frac{\mathcal{V}_{-,j}}{p^{est}_{-,j}} \right) e_1\end{aligned}$$

where

$$\mathcal{V}_{\pm,j} = \frac{\hat{M}^{-1}_{\pm,j}}{\hat{\tau}^t_j(Z_i; \Pi, \mathcal{S}^{te})^2} \left(\hat{\sigma}^{2,y}_{\pm,j} + \frac{\hat{\tau}^y_j(Z_i; \Pi, \mathcal{S}^{te})^2}{\hat{\tau}^t_j(Z_i; \Pi, \mathcal{S}^{te})} \hat{\sigma}^{2,t}_{\pm,j} + \frac{\hat{\tau}^y_j(Z_i; \Pi, \mathcal{S}^{te})}{\hat{\tau}^t_j(Z_i; \Pi, \mathcal{S}^{te})} \hat{C}^{y,t}_{\pm,j} \right)$$

is the within leaf variance of the outcome equation at the threshold, estimated from above (+) or below (-) and $\hat{\tau}^t_j(\cdot), \hat{\tau}^y_j(\cdot)$ are the j'th leaf treatment effect estimated on the participation equation ($\hat{\tau}^t_j(\cdot)$) and on the outcome equation ($\hat{\tau}^y_j(\cdot)$). $\sigma^{2,t}_{\pm,j}$, $\sigma^{2,y}_{\pm,j}$ and $C^{y,t}_{\pm,j}$ are estimators for the variances and co-variance for the leaf-by-leaf disturbance terms. See the derivations in the Appendix, Section C.

The EMSE criterion for the fuzzy design combines the jumps in the outcome and in the participation equation along with the variance. This means that if there is a difference in two groups in the participation probabilities at the threshold or in the outcome equation, then the EMSE criterion results in a lower value and finds this difference. Similarly, if the variance of $\hat{\tau}^y$, $\hat{\tau}^t$ or their co-variance gets lower by a split, the EMSE criterion will be lower, thus even if there is no big change in the treatment effect, but there is in its variance, the algorithm considers this split.

One feature of this criterion is that, if the changes in the jump in the outcome equation and in the participation equation are with the same magnitude – resulting in the same treatment effect – then the EMSE criterion does not change. If one is interested in heterogeneity in the participation effect and in the intent-to-treat effect separately as well, then it is possible to use a sharp design for both equations separately and then assemble the results from the two trees.

7 Conclusions

The paper proposes an algorithm, that uncovers treatment effect heterogeneity in classical regression discontinuity (RD) designs. Heterogeneity is identified through the values of pre-treatment covariates and it is the task of the algorithm to find the relevant groups. The introduced honest "*regression discontinuity tree*" algorithm ensures a fairly flexible functional form for the conditional treatment effect (CATE) with valid inference while handling many potential pre-treatment covariates and their interactions.

The properties of the CATE function for sharp regression design are analyzed and the paper shows how the algorithm works. An estimable EMSE criterion is derived, which uses the specifics of RD setup, such as the distinct estimation of polynomial functions below and above the cutoff value. Furthermore, the algorithm utilizes the honest approach to get a valid inference for the parameter of interest.

Monte Carlo simulation results show that the proposed algorithm and criterion work well and discover the true tree in more than 95% of the cases. The estimated conditional treatment effects are unbiased and the standard errors provide proper estimates for 95% confident interval coverage.

Finally, the paper shows how one can utilize the algorithm in practice. I use Pop-Eleches and Urquiola (2013) data on the Romanian school system and uncover heterogeneous treatment effects on the impact of going to a better school. The algorithm shows a more detailed picture when revisiting the heterogeneity analysis done by Pop-Eleches and Urquiola (2013). The results suggest, i) in the most competitive schools without an outside option, students are less likely to take the Baccalaureate exam when going to a better school, indicating a negative peer effect. ii) there is no positive peer quality effect for students who scored in the lowest 35% with internet access. The discovery of these groups encourages further investigations and this future research may help to understand better the allocation mechanism of the Romanian school system.

References

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S. and Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *Stat*, 1050(5).

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.

Bargagli, S. and Gnecco, G. (2020). Causal tree with instrumental variable: an extension of the causal tree framework to irregular assignment mechanisms. *International Journal of Data Science and Analytics*, 9:315–337.

Becker, S. O., Egger, P. H., and von Ehrlich, M. (2013). Absorptive capacity and the growth and investment effects of regional transfers: A regression discontinuity design with heterogeneous treatment effects. *American Economic Journal: Economic Policy*, 5(4):29–77.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Caetano, C., Caetano, G., and Escanciano, J. C. (2017). Over-identified regression discontinuity design. *Unpublished, University of Rochester*.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 101(3):442–451.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G., and Keele, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4):1229–1248.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):1–68.

Fan, Q., Hsu, Y.-C., Lieli, R. P., and Zhang, Y. (2020). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business and Economic Statistics*. forthcoming.

Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, pages 1–15.

Gelman, A. and Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics*, 37(3):447–456.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer.

Hsu, Y.-C. and Shen, S. (2019). Testing treatment effect heterogeneity in regression discontinuity designs. *Journal of Econometrics*, 208(2):468–486.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Knaus, M., Lechner, M., and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *Econometrics Journal*, 24(1):134–161.

Knaus, M. C. (2021). Double machine learning based program evaluation under unconfoundedness. Working paper, `https://arxiv.org/abs/2003.03191`.

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

Ludwig, J. and Miller, D. L. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, 122(1):159–208.

Nekipelov, D., Novosad, P., and Ryan, S. P. (2019). Moment forests. Working paper, `https://cpb-us-w2.wpmucdn.com/sites.wustl.edu/dist/5/501/files/2016/10/momentTrees.pdf`.

Pei, Z., Lee, D. S., Card, D., and Weber, A. (2020). Local polynomial order in regression discontinuity designs. Working Paper 27424, National Bureau of Economic Research.

Pop-Eleches, C. and Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, 103(4):1289–1324.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge university press.

Robson, M., Doran, T., Cookson, R., et al. (2019). Estimating and decomposing conditional average treatment effects: the smoking ban in england. Technical report, HEDG, Department of Economics, University of York.

Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, 78(1):169–211.

Semenova, V. and Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.

Toda, T., Wakano, A., and Hoshino, T. (2019). Regression discontinuity design with multiple groups for heterogeneous causal effect estimation. Technical report, arXiv.

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

Xu, K.-L. (2017). Regression discontinuity with categorical outcomes. *Journal of Econometrics*, 201(1):1–18.

Appendix

A Decomposition of EMSE criterion

Here I provide the decomposition of $EMSE_\tau(\Pi)$ criterion.

$$
\begin{aligned}
EMSE_\tau(\Pi) &= \mathbb{E}_{\mathcal{S}^{te},\mathcal{S}^{est}}\left[MSE_\tau(\mathcal{S}^{te},\mathcal{S}^{est},\Pi)\right] \\
&= \mathbb{E}_{X_i,Z_i,\mathcal{S}^{est}}\left\{\left[\tau(Z_i)-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right]^2-\tau^2(Z_i)\right\} \\
&= \mathbb{E}_{X_i,Z_i,\mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{est})-2\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\tau(Z_i)\right\} \\
&\text{using law of iterated expectations} \\
&= \mathbb{E}_{X_i,Z_i,\mathcal{S}^{est}}\left\{\mathbb{E}\left[\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{est})-2\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\tau(Z_i)\mid X_i=c,\mathbb{1}_{\ell_1}(Z_i),\ldots,\mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right]\right\} \\
&= \mathbb{E}_{X_i,Z_i,\mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{est})-2\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\mathbb{E}\left[\tau(Z_i)\mid X_i=c,\mathbb{1}_{\ell_1}(Z_i),\ldots,\mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right]\right\} \\
&\text{as } \mathbb{E}\left[\tau(Z_i)\mid X_i=c,\mathbb{1}_{\ell_1}(Z_i),\ldots,\mathbb{1}_{\ell_{\#\Pi}}(Z_i)\right]=\tau(Z_i;\Pi) \\
&= \mathbb{E}_{Z_i,\mathcal{S}^{est}}\left\{\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{est})-2\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\tau(Z_i;\Pi)\right\} \\
&= \mathbb{E}_{Z_i,\mathcal{S}^{est}}\left\{\left[\tau(Z_i;\Pi)-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right]^2-\tau^2(Z_i;\Pi)\right\} \\
&\text{using } Z_i \perp\!\!\!\perp \mathcal{S}^{est} \\
&= \mathbb{E}_{Z_i,\mathcal{S}^{est}}\left\{\left[\tau(Z_i;\Pi)-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right]^2\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
&\text{using law of iterated expectations and } Z_i \perp\!\!\!\perp \mathcal{S}^{est} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\tau(Z_i;\Pi)-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right)^2\mid Z_i\right]\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
&\text{Note: } \mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]=\tau(z;\Pi) \text{ where } z \text{ is fixed, thus} \\
&\tau(Z_i;\Pi)=\mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\overset{Z_i \perp\!\!\!\perp \mathcal{S}^{est}}{=}\mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\mid Z_i\right] \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{est}}\left[\left(\mathbb{E}_{\mathcal{S}^{est}}\left[\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\mid Z_i\right]-\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\right)^2\mid Z_i\right]\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(Z_i;\Pi,\mathcal{S}^{est})\mid Z_i\right]\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\}-\mathbb{E}_{Z_i}\left\{\tau^2(Z_i;\Pi)\right\} \qquad \blacksquare
\end{aligned}
$$

B Derivation of honest sharp RDD criterion

In the following I derive the estimators for the EMSE function for regression discontinuity tree.

$$EMSE_\tau(\Pi)=\mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\}-\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right]$$

Let me consider the two parts separately, starting with the expected variance, then the expected square term and finally I put them together.

B.1 Expected variance of CATE

Let start with the expected variance part and focus on the variance itself. Here z is fixed, thus

$$\begin{aligned}
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_{+}(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_{-}(c,z;\Pi,\mathcal{S}^{est})\right] \\
&= \mathbb{V}_{\mathcal{S}^{est}}\left[e_1' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\,\hat{\boldsymbol{\delta}}_j^{+,est}\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[e_1' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\,\hat{\boldsymbol{\delta}}_j^{-,est}\right] \\
&= \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\,\mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\hat{\boldsymbol{\delta}}_j^{+,est}\right] + \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\,\mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\hat{\boldsymbol{\delta}}_j^{-,est}\right] \\
&= \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\,\left(e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,est}\right] e_1\right) + \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\,\left(e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,est}\right] e_1\right) \\
&= \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)\,e_1'\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,est}\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,est}\right]\right) e_1
\end{aligned}$$

where $e_1 = [1, 0 \ldots, 0]$ is still a $1 \times (p+1)$ selector-vector. Because $\mathcal{S}^{est} \perp\!\!\!\perp \mathcal{S}^{te}$, $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,est}\right]$ and $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,est}\right]$ can be estimated using the test sample and the additional knowledge for the number of observations in the estimation sample to adjust for sample size. In the case of homoscedastic disturbance term within each leaf, the estimator for the variances are

$$\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,est}\right] = \frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}}\,, \qquad \hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,est}\right] = \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}$$

where

$$\begin{aligned}
N^{est}_{+,j} &= \sum_{i\in\mathcal{S}^{est}} \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\,, \qquad N^{te}_{+,j} = \sum_{i\in\mathcal{S}^{te}} \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \\
\hat{M}_{+,j} &= \frac{1}{N^{te}_{+,j}} \sum_{i\in\mathcal{S}^{te}} \boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \\
\hat{\sigma}^2_{+,j} &= \frac{1}{N^{te}_{+,j} - q - 1} \sum_{i\in\mathcal{S}^{te}} \left(Y_i - \boldsymbol{X}_i'\hat{\boldsymbol{\delta}}_j^{+,te}\right)^2 \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\,, \\
j &= 1, 2, \ldots, \#\Pi\,.
\end{aligned}$$

Same applies for the components of $\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{-,est}\right]$, but using observations, below the threshold, selected by $1 - \mathbb{1}_c(X_i)$ instead of $\mathbb{1}_c(X_i)$.

Using these estimates, leads to the following expression for the variance, with scalar z,

$$\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right] = \sum_{j=1}^{\#\Pi}\left\{\mathbb{1}_{\ell_j}(z;\Pi)\; e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right]\right\} e_1$$

Now, I can express the expected value of this expression over the features in the test sample. A natural estimator is the mean of the variances, using Z_i values from the test sample,

$$\begin{aligned}
\hat{\mathbb{E}}_{Z_i}\left\{\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} &= \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\left\{\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(Z_i;\Pi)e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \sum_{j=1}^{\#\Pi}\left\{\left(\frac{\sum_{i\in\mathcal{S}^{te}}\mathbb{1}_{\ell_j}(Z_i;\Pi)}{N^{te}}\right) e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \sum_{j=1}^{\#\Pi}\left\{\frac{N^{te}_j}{N^{te}}e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{N^{est}\frac{N^{te}_j}{N^{te}}e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{\left(\frac{N^{te}_j}{N^{te}}N^{est}\frac{N^{est}_j}{N^{est}_j}\right) e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}}\right] e_1\right\} \\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{\left(\frac{N^{te}_j}{N^{te}}\frac{N^{est}}{N^{est}_j}\right) e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}/N^{est}_j} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}/N^{est}_j}\right] e_1\right\} \\
&\quad \text{using assumption for same leaf shares: } \frac{N^{te}_j}{N^{te}} \approx \frac{N^{est}_j}{N^{est}} \\
&\approx \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{N^{est}_{+,j}/N^{est}_j} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{N^{est}_{-,j}/N^{est}_j}\right] e_1\right\} \\
&= \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{\left[e_1'\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right] e_1\right\}
\end{aligned}$$

where, N^{te}_j, N^{est}_j are the number of observations within leaf j for the test sample and estimation sample, respectively, and $p^{est}_{\pm,j}$ is the share of units above (+) and below (-) the threshold. This derivation uses the fact that observations are randomly assigned to the test sample and to the estimation sample, thus the leaf shares in the test sample (N^{te}_j/N^{te}) is approximately the same as in the estimation sample, (N^{est}_j/N^{est}).

B.2 Expected square of CATE

The second part of the EMSE criterion is the estimator for the expected squared of the true CATE, $\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right]$ over the test sample's features. Using, $\tau(z;\Pi) = \mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]$, where z is fixed, therefore $\tau(Z_i;\Pi) = \mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}$. Based on this fact, it follows:

$$\begin{aligned}
\mathbb{E}_{Z_i}\left[\tau^2(Z_i;\Pi)\right] &= \mathbb{E}_{Z_i}\left\{\mathbb{E}_{\mathcal{S}^{te}}\left[\hat{\tau}^2(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\} \\
&\quad \text{using variance decomposition} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}^2_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i} - \mathbb{V}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\} \\
&= \mathbb{E}_{Z_i}\left\{\mathbb{E}^2_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\}
\end{aligned}$$

The two parts can be estimated by two natural candidates. The expected square CATE is just the average of the squared CATE estimator given by the test sample. The expected variance term is similar to the previous, but note that the variance is estimated purely on the test sample. This means that the scaling factor for the number of observations is coming only from the test sample.

$$\begin{aligned}
\hat{\mathbb{E}}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{te}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{te})\right]\Big|_{z=Z_i}\right\} &= \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{te}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{te}_{-,j}}\right]e_1\right\} \\
&\quad \text{using assumption for same obs. shares within each leaf:} \\
&\quad p^{te}_{+,j} \approx p^{est}_{+,j},\ p^{te}_{-,j} \approx p^{est}_{-,j},\ \forall j \\
&= \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\}
\end{aligned}$$

This expression is the same as the expected variance using the test sample, the only difference is the scalar N^{te} is used instead of N^{est}. The assumption for the same observation shares is used here to make the weights the same for the variance estimators.

The estimator for the expected value of the true squared CATE function over the test sample is given by,

$$\hat{\mathbb{E}}_{Z_i}\left[\tau^2(Z_i;\Pi)\right] = \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}^2(Z_i;\Pi,\mathcal{S}^{te}) - \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left[\frac{\hat{\sigma}^2_{+,j}\hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j}\hat{M}^{-1}_{-,j}}{p^{est}_{-,j}}\right]e_1\right\}$$

B.3 Estimator for EMSE

Plugging the two parts together yields an estimator for the EMSE criterion,

$$\widehat{EMSE}_{\tau}(\mathcal{S}^{te}, \mathcal{S}^{est}, \Pi) = -\frac{1}{N^{te}} \sum_{i \in \mathcal{S}^{te}} \hat{\tau}^2(Z_i; \Pi, \mathcal{S}^{te}) + \left(\frac{1}{N^{te}} + \frac{1}{N^{est}}\right) \sum_{j=1}^{\#\Pi} \left\{ e_1' \left[\frac{\hat{\sigma}^2_{+,j} \hat{M}^{-1}_{+,j}}{p^{est}_{+,j}} + \frac{\hat{\sigma}^2_{-,j} \hat{M}^{-1}_{-,j}}{p^{est}_{-,j}} \right] e_1 \right\}$$

C Derivation of honest fuzzy RDD leaf-by-leaf LS criterion

Let assume, that there is a sample $\mathcal{S}$, $i = 1, \ldots, N$ with identically and independently distributed observations of (Y_i, X_i, T_i, Z_i). For leaf-by-leaf estimation, I use the fact, $\mathbb{1}_{\ell_j}(z; \Pi)$ creates disjoint sets, and one can estimate the parameters and their variances consistently in each leaf separately. The conditional mean estimator is given by

$$\hat{\mu}^t_+(x, z; \Pi, \mathcal{S}) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \boldsymbol{\delta}^{+,t}_j \quad , \quad \hat{\mu}^t_-(x, z; \Pi, \mathcal{S}) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \boldsymbol{\delta}^{-,t}_j$$
$$\hat{\mu}^y_+(x, z; \Pi, \mathcal{S}) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \boldsymbol{\delta}^{+,y}_j \quad , \quad \hat{\mu}^y_-(x, z; \Pi, \mathcal{S}) = \boldsymbol{X}' \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \boldsymbol{\delta}^{-,y}_j$$

where $\boldsymbol{\delta}^{+,t}_j, \boldsymbol{\delta}^{-,t}_j, \boldsymbol{\delta}^{+,y}_j$ and $\boldsymbol{\delta}^{-,y}_j$ estimated by OLS:

$$\hat{\boldsymbol{\delta}}^{+,t}_j = \arg\min_{\boldsymbol{\delta}^{+,t}_j} \sum_{i \in \mathcal{S}} \mathbb{1}_c(x) \mathbb{1}_{\ell_j}(z; \Pi) \left(T_i - \boldsymbol{X}_i' \boldsymbol{\delta}^{+,t}_j\right)^2$$
$$\hat{\boldsymbol{\delta}}^{-,t}_j = \arg\min_{\boldsymbol{\delta}^{-,t}_j} \sum_{i \in \mathcal{S}} (1 - \mathbb{1}_c(x)) \mathbb{1}_{\ell_j}(z; \Pi) \left(T_i - \boldsymbol{X}_i' \boldsymbol{\delta}^{-,t}_j\right)^2$$
$$\hat{\boldsymbol{\delta}}^{+,y}_j = \arg\min_{\boldsymbol{\delta}^{+,y}_j} \sum_{i \in \mathcal{S}} \mathbb{1}_c(x) \mathbb{1}_{\ell_j}(z; \Pi) \left(Y_i - \boldsymbol{X}_i' \boldsymbol{\delta}^{+,y}_j\right)^2$$
$$\hat{\boldsymbol{\delta}}^{-,y}_j = \arg\min_{\boldsymbol{\delta}^{-,y}_j} \sum_{i \in \mathcal{S}} (1 - \mathbb{1}_c(x)) \mathbb{1}_{\ell_j}(z; \Pi) \left(Y_i - \boldsymbol{X}_i' \boldsymbol{\delta}^{-,y}_j\right)^2$$

Estimator for CLATE parameter based on these polynomial functions is given by

$$\hat{\tau}_{FRD}(z; \Pi, \mathcal{S}) = \frac{\hat{\mu}^y_+(c, z; \Pi, \mathcal{S}) - \hat{\mu}^y_-(c, z; \Pi, \mathcal{S})}{\hat{\mu}^t_+(c, z; \Pi, \mathcal{S}) - \hat{\mu}^t_-(c, z; \Pi, \mathcal{S})} = \frac{\hat{\tau}^y(z; \Pi, \mathcal{S})}{\hat{\tau}^t(z; \Pi, \mathcal{S})} = \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z; \Pi) \frac{\hat{\alpha}^y_{+,j} - \hat{\alpha}^y_{-,j}}{\hat{\alpha}^t_{+,j} - \hat{\alpha}^t_{-,j}}$$

and its variance:

$$\begin{aligned}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S})\right] = &\frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S})^2}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S})\right] \\ &+ \frac{\hat{\tau}^y(z;\Pi,\mathcal{S})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S})^4}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^t(z;\Pi,\mathcal{S})\right] \\ &- 2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S})}{\hat{\tau}^t(z;\Pi,\mathcal{S})^3}\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S}),\hat{\tau}^t(z;\Pi,\mathcal{S})\right]\end{aligned}$$

where $\mathbb{C}_{\mathcal{S}^{est}}[\cdot,\cdot]$ is the covariance of two random variable. Each part can be decomposed one step further,

$$\begin{aligned}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S})\right] \\ \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}^t(z;\Pi,\mathcal{S})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S})\right] \\ \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\tau}^y(z;\Pi,\mathcal{S}),\hat{\tau}^t(z;\Pi,\mathcal{S})\right] &= \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}),\hat{\mu}^t_+(c,z;\Pi,\mathcal{S})\right] \\ &\quad + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}),\hat{\mu}^t_-(c,z;\Pi,\mathcal{S})\right]\end{aligned}$$

I use the same expected MSE criterion for fuzzy design as well. After the same manipulations as in Section A, one gets:

$$EMSE_\tau(\Pi) = \mathbb{E}_{Z_i}\left\{\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} - \mathbb{E}_{Z_i}\left[\tau^2_{FRD}(Z_i;\Pi)\right]$$

One can construct estimators for these two terms. The variance part from the expected variance is

$$\begin{aligned}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}(z;\Pi,\mathcal{S}^{est})\right] = &\frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est})\right]\right) \\ &+ \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4}\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right]\right) \\ &- 2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3}(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] \\ &\quad + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right])\end{aligned}$$

Decomposing $\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est})\right]$:

$$\begin{aligned}\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\hat{\boldsymbol{\delta}}_j^{+,y,est}\right]\\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\mathbb{V}_{\mathcal{S}^{est}}\left[e_1'\hat{\boldsymbol{\delta}}_j^{+,y,est}\right]\\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est}\right]e_1\end{aligned}$$

and $\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}_+^t(c,z;\Pi,\mathcal{S}^{est})\right]$:

$$\begin{aligned}\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}_+^y(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}_+^t(c,z;\Pi,\mathcal{S}^{est})\right] &= \mathbb{C}_{\mathcal{S}^{est}}\left[e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\hat{\boldsymbol{\delta}}_j^{+,y,est},e_1'\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\hat{\boldsymbol{\delta}}_j^{+,t,est}\right]\\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)\mathbb{C}_{\mathcal{S}^{est}}\left[e_1'\hat{\boldsymbol{\delta}}_j^{+,y,est},e_1'\hat{\boldsymbol{\delta}}_j^{+,t,est}\right]\\ &= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est},\hat{\boldsymbol{\delta}}_j^{+,t,est}\right]e_1\end{aligned}$$

All the other variances/covariance have the same form with the appropriate parameter vector. Because $\mathcal{S}^{est} \perp\!\!\!\perp \mathcal{S}^{te}$, one can estimate all the variances and covariances using the observations from the test sample and use only the additional knowledge on the number of observations in the estimation sample. In the simplest – finite variances of the error terms within each leaf – one can write the following sample analogs (below threshold units it is similar).

$$\widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est}\right] = \frac{\hat{\sigma}_{+,j}^{2,y}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}, \quad \widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\boldsymbol{\delta}}_j^{+,t,est}\right] = \frac{\hat{\sigma}_{+,j}^{2,t}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}, \quad \widehat{\mathbb{C}_{\mathcal{S}^{est}}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est},\hat{\boldsymbol{\delta}}_j^{+,t,est}\right] = \frac{\hat{C}_{+,j}^{y,t}\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}$$

where

$$N^{est}_{+,j} = \sum_{i \in \mathcal{S}^{est}} \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \quad , \quad N^{te}_{+,j} = \sum_{i \in \mathcal{S}^{te}} \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)$$

$$\hat{M}_{+,j} = \frac{1}{N^{te}_{+,j}} \sum_{i \in \mathcal{S}^{te}} \boldsymbol{X}_i \boldsymbol{X}_i' \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)$$

$$\hat{\sigma}^{2,y}_{+,j} = \frac{1}{N^{te}_{+,j} - p - 1} \sum_{i \in \mathcal{S}^{te}} \left[(\epsilon^y_i)^2 \, \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \right] \ , \qquad \epsilon^y_i = Y_i - \boldsymbol{X}_i' \hat{\boldsymbol{\delta}}^{+,y,te}_j$$

$$\hat{\sigma}^{2,t}_{+,j} = \frac{1}{N^{te}_{+,j} - p - 1} \sum_{i \in \mathcal{S}^{te}} \left[(\epsilon^t_i)^2 \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \right] \ , \qquad \epsilon^t_i = T_i - \boldsymbol{X}_i' \hat{\boldsymbol{\delta}}^{+,t,te}_j$$

$$\hat{C}^{y,t}_{+,j} = \frac{1}{N^{te}_{+,j} - p - 1} \sum_{i \in \mathcal{S}^{te}} \left(\epsilon^y_i \epsilon^t_i \mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i) \right) \ , \qquad j = 1, 2, \ldots, \#\Pi$$

Remark: the number of observations and the inverse of the running variable's product is the same for both treatment and outcome equations. It is also easy to use other variance estimators (e.g., heteroscedastic-robust versions or clustered), see Appendix D.
Putting together the variances, in the homoscedastic case I have the following expression,

$$\begin{aligned}
\widehat{\mathbb{V}_{\mathcal{S}^{est}}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S}^{est})\right] &= \frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2} \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)(z) \left[\frac{e_1' \left(\hat{\sigma}^{2,y}_{+,j} \hat{M}^{-1}_{+,j} \right) e_1}{N^{est}_{+,j}} + \frac{e_1' \left(\hat{\sigma}^{2,y}_{-,j} \hat{M}^{-1}_{-,j} \right) e_1}{N^{est}_{-,j}} \right] \\
&+ \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4} \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)(z) \left[\frac{e_1' \left(\hat{\sigma}^{2,t}_{+,j} \hat{M}^{-1}_{+,j} \right) e_1}{N^{est}_{+,j}} + \frac{e_1' \left(\hat{\sigma}^{2,t}_{-,j} \hat{M}^{-1}_{-,j} \right) e_1}{N^{est}_{-,j}} \right] \\
&- 2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3} \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)(z) \left[\frac{e_1' \left(\hat{C}^{y,t}_{+,j} \hat{M}^{-1}_{+,j} \right) e_1}{N^{est}_{+,j}} + \frac{e_1' \left(\hat{C}^{y,t}_{-,j} \hat{M}^{-1}_{-,j} \right) e_1}{N^{est}_{-,j}} \right] \\
&= \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)(z) e_1' \left(\frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2} \left[\frac{\left(\hat{\sigma}^{2,y}_{+,j} \hat{M}^{-1}_{+,j} \right)}{N^{est}_{+,j}} + \frac{\left(\hat{\sigma}^{2,y}_{-,j} \hat{M}^{-1}_{-,j} \right)}{N^{est}_{-,j}} \right] \right. \\
&\quad + \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4} \left[\frac{\left(\hat{\sigma}^{2,t}_{+,j} \hat{M}^{-1}_{+,j} \right)}{N^{est}_{+,j}} + \frac{\left(\hat{\sigma}^{2,t}_{-,j} \hat{M}^{-1}_{-,j} \right)}{N^{est}_{-,j}} \right] \\
&\quad \left. - 2\frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3} \left[\frac{\left(\hat{C}^{y,t}_{+,j} \hat{M}^{-1}_{+,j} \right)}{N^{est}_{+,j}} + \frac{\left(\hat{C}^{y,t}_{-,j} \hat{M}^{-1}_{-,j} \right)}{N^{est}_{-,j}} \right] \right) e_1 \\
&= \sum_{j=1}^{\#\Pi} \mathbb{1}_{\ell_j}(z;\Pi)(z) e_1' \left(\frac{\mathcal{V}_{+,j}}{N^{est}_{+,j}} + \frac{\mathcal{V}_{-,j}}{N^{est}_{-,j}} \right) e_1
\end{aligned}$$

where

$$\mathcal{V}_{+,j} = \frac{\hat{M}_{+,j}^{-1}}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\left(\hat{\sigma}_{+,j}^{2,y} + \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\hat{\sigma}_{+,j}^{2,t} + \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})}\hat{C}_{+,j}^{y,t}\right)$$
$$\mathcal{V}_{-,j} = \frac{\hat{M}_{-,j}^{-1}}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\left(\hat{\sigma}_{-,j}^{2,y} + \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}\hat{\sigma}_{-,j}^{2,t} + \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})}\hat{C}_{-,j}^{y,t}\right)$$

The expected value of this variance over Z_i from the test sample can be calculated similarly as in the sharp RDD case.

$$\begin{aligned}\hat{\mathbb{E}}_{Z_i}\left\{\hat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\tau}_{FRD}(z;\Pi,\mathcal{S}^{est})\right]\Big|_{z=Z_i}\right\} &= \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\left\{\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)(z)e_1'\left(\frac{\mathcal{V}_{+,j}}{N_{+,j}^{est}} + \frac{\mathcal{V}_{-,j}}{N_{-,j}^{est}}\right)e_1\right\} \\ &\approx \frac{1}{N^{est}}\sum_{j=1}^{\#\Pi}\left\{e_1'\left(\frac{\mathcal{V}_{+,j}}{p_{+,j}^{est}} + \frac{\mathcal{V}_{-,j}}{p_{-,j}^{est}}\right)e_1\right\}\end{aligned}$$

The second part of the EMSE criterion is the estimator for the expected squared $\tau_{FRD}^2(Z_i;\Pi)$. Similarly to sharp RD, one can construct the following estimator,

$$\hat{\mathbb{E}}_{Z_i}\left[\tau_{FRD}^2(Z_i;\Pi)\right] = \frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}_{FRD}^2(Z_i;\Pi,\mathcal{S}^{te}) - \frac{1}{N^{te}}\sum_{j=1}^{\#\Pi}e_1'\left(\frac{\mathcal{V}_{+,j}}{p_{+,j}^{est}} + \frac{\mathcal{V}_{-,j}}{p_{-,j}^{est}}\right)e_1$$

Note, here everything is estimated on the test sample and I used the assumption, that the number of unit shares for below and above the threshold – for all leaf – are approximately the same in the estimation and test sample ($p_{+,j}^{te} \approx p_{+,j}^{est}, p_{-,j}^{te} \approx p_{-,j}^{est}$). The feasible criteria for fuzzy design for EMSE:

$$\begin{aligned}\widehat{EMSE}_{\tau_{FRD}}(\mathcal{S}^{te},\mathcal{S}^{est},\Pi) = &-\frac{1}{N^{te}}\sum_{i\in\mathcal{S}^{te}}\hat{\tau}_{FRD}^2(Z_i;\Pi,\mathcal{S}^{te}) \\ &+\left(\frac{1}{N^{te}} + \frac{1}{N^{est}}\right)\sum_{j=1}^{\#\Pi}e_1'\left(\frac{\mathcal{V}_{+,j}}{p_{+,j}^{est}} + \frac{\mathcal{V}_{-,j}}{p_{-,j}^{est}}\right)e_1\end{aligned}$$

D Derivation of variances for leaf-by-leaf LS criterion

Homoscedastic error assumption is rather a strong assumption in RD context, thus the use of different heteroscedastic consistent estimators is favorable. First, I show derivation of $\widehat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est}\right]$ – the other parts can be calculated similarly – then I put together with the other parts.

General case:

$$\begin{aligned}\widehat{\mathbb{V}}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}_j^{+,y,est}\right] &= \frac{1}{N_{+,j}^{est}}\left(\frac{1}{N_{+,j}^{te}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right)^{-1}\left[\frac{1}{N_{+,j}^{te}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\hat{\Omega}\boldsymbol{X}_i\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right] \\ &\qquad\left(\frac{1}{N_{+,j}^{te}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i\boldsymbol{X}_i'\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right)^{-1} \\ &= \frac{1}{N_{+,j}^{est}}\hat{M}_{+,j}^{-1}\left[\frac{1}{N_{+,j}^{te}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\hat{\Omega}\boldsymbol{X}_i\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)\right]\hat{M}_{+,j}^{-1} \\ &= \frac{1}{N_{+,j}^{est}}\hat{M}_{+,j}^{-1}\hat{\Sigma}_{+,j}\hat{M}_{+,j}^{-1}\end{aligned}$$

Estimators are different in how to calculate $\hat{\Sigma}_{+,j}$:
White's estimator ('HCE0'):

$$\hat{\Sigma}_{+,j}^{HCE0} = \frac{1}{N_{+,j}^{te}}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\boldsymbol{X}_i(\epsilon_i^y)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)$$

Adjusted 'HCE1':

$$\hat{\Sigma}_{+,j}^{HCE1} = \frac{1}{N_{+,j}^{te}-p-1}\sum_{i\in\mathcal{S}^{te}}\boldsymbol{X}_i'\boldsymbol{X}_i(\epsilon_i^y)^2\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)$$

In case of clustered SE, with HC1

$$\hat{\Sigma}_{+,j}^{C} = \frac{N_{+,j}^{te}-1}{(N_{+,j}^{te}-p-1)^2}\frac{G_{+,j}^{te}}{G_{+,j}^{te}-1}\sum_{i\in\mathcal{S}^{te}}\left(\sum_{c=1}^{G_{+,j}^{te}}\boldsymbol{X}_{i,c}'\boldsymbol{X}_{i,c}(\epsilon_{i,c}^y)^2\right)\mathbb{1}_{\ell_j}(Z_i;\Pi)\mathbb{1}_c(X_i)$$

where $G_{+,j}^{te}$ is the number of clusters in leaf j above the threshold in the test sample. The variance estimators are similarly constructed for parameters below the threshold.

In sharp RD, one gets the variance estimator as,

$$\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_{SRD}(z;\Pi,\mathcal{S}^{est})\right] = \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(Z_i;\Pi)e_1'\left\{\frac{\hat{M}_{+,j}^{-1}\hat{\Sigma}_+\hat{M}_{+,j}^{-1}}{N_{+,j}^{est}}+\frac{\hat{M}_{-,j}^{-1}\hat{\Sigma}_-\hat{M}_{-,j}^{-1}}{N_{-,j}^{est}}\right\}e_1$$

In fuzzy RD, let $A_1 = \frac{1}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^2}$, $A_2 = \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})^2}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^4}$ and $A_3 = \frac{\hat{\tau}^y(z;\Pi,\mathcal{S}^{est})}{\hat{\tau}^t(z;\Pi,\mathcal{S}^{est})^3}$. Putting together

the variance for CLATE parameters,

$$
\begin{aligned}
\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\tau}_F(z;\Pi,\mathcal{S}^{est})\right] &= \left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est})\right]\right) \\
&+ A_2\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right]\right) \\
&- 2A_3(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_+(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_+(c,z;\Pi,\mathcal{S}^{est})\right] \\
&\qquad + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\mu}^y_-(c,z;\Pi,\mathcal{S}^{est}),\hat{\mu}^t_-(c,z;\Pi,\mathcal{S}^{est})\right]) \\
&= A_1\left(\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,y,est}_j\right]e_1 + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,y,est}_j\right]e_1\right) \\
&+ A_2\left(\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,t,est}_j\right]e_1 + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,t,est}_j\right]e_1\right) \\
&- 2A_3\left(\sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,y,est}_j,\hat{\boldsymbol{\delta}}^{+,t,est}_j\right]e_1\right. \\
&\qquad \left. + \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,y,est}_j,\hat{\boldsymbol{\delta}}^{-,t,est}_j\right]e_1\right) \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{A_1\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,y,est}_j\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,y,est}_j\right]\right)\right. \\
&\qquad + A_2\left(\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,t,est}_j\right] + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,t,est}_j\right]\right) \\
&\qquad \left. -2A_3\left(\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,y,est}_j,\hat{\boldsymbol{\delta}}^{+,t,est}_j\right] + \mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,y,est}_j,\hat{\boldsymbol{\delta}}^{-,t,est}_j\right]\right)\right\}e_1 \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{A_1\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,y,est}_j\right] + A_2\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,t,est}_j\right] - 2A_3\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{+,y,est}_j,\hat{\boldsymbol{\delta}}^{+,t,est}_j\right]\right. \\
&\qquad \left. +A_1\mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,y,est}_j\right] + A_2 + \mathbb{V}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,t,est}_j\right] - 2A_3\mathbb{C}_{\mathcal{S}^{est}}\left[\hat{\boldsymbol{\delta}}^{-,y,est}_j,\hat{\boldsymbol{\delta}}^{-,t,est}_j\right]\right\}e_1 \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{\frac{1}{N^{est}_{+,j}}\hat{M}^{-1}_{+,j}\left(A_1\hat{\Sigma}^y_{+,j} + A_2\hat{\Sigma}^t_{+,j} - 2A_3\hat{C}_{+,j}\right)\hat{M}^{-1}_{+,j}\right. \\
&\qquad \left. +\frac{1}{N^{est}_{-,j}}\hat{M}^{-1}_{-,j}\left(A_1\hat{\Sigma}^y_{-,j} + A_2\hat{\Sigma}^t_{-,j} - 2A_3\hat{C}_{-,j}\right)\hat{M}^{-1}_{-,j}\right\}e_1 \\
&= \sum_{j=1}^{\#\Pi}\mathbb{1}_{\ell_j}(z;\Pi)e_1'\left\{\frac{1}{N^{est}_{+,j}}\hat{M}^{-1}_{+,j}\hat{\Sigma}^*_+\hat{M}^{-1}_{+,j} + \frac{1}{N^{est}_{-,j}}\hat{M}^{-1}_{-,j}\hat{\Sigma}^*_-\hat{M}^{-1}_{-,j}\right\}e_1
\end{aligned}
$$

This result is quite useful: there is no need to calculate and multiply with $\hat{M}^{-1}_{\pm,j}$ multiple times during calculating the variances, but they can be 'added up', using only the test sample.

E Monte Carlo simulation setup

For Monte Carlo simulations, I use a general formulation for the DGPs and change the appropriate parts for each specific setup.

$$Y_i = \eta(X_i, Z_{i,k}) + \mathbb{1}_c(X_i) \times \kappa(Z_{i,k}) + \epsilon_i$$

where $\eta(X_i, Z_{i,k})$ is the conditional expectation function, which is depending on the running variable (X_i) and can be a function of the features ($Z_{i,k}$) as well. The disturbance term is generated from a normal distribution $\epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon)$. I generate $k = 1, \ldots, K$ features such that $Z_{i,k}$ is independent across k and independent from ϵ_i, X_i. The source of variation comes from ϵ_i during the simulations, thus $X_i, Z_{i,k}$ are the same across the Monte Carlo samples. All the other terms are dependent on the setup.
I report three Monte Carlo average statistics to evaluate the performance of the algorithm:

1. Average of the infeasible MSE: $MSE = \frac{1}{N^{eval}} \sum_{i=1}^{N^{eval}} (\kappa(Z_{i,k}) - \hat{\tau}(Z_i; \hat{\Pi}(\mathcal{S}^{tr}), \mathcal{S}^{est}))^2$
2. Average number of leaves in the final tree.
3. DGP found: this is only feasible for DGPs, where the DGP itself has a tree structure. The DGP is said to be found in the used features for the final tree is the same as for the DGP. [45]

For DGP 1 and 2, I use linear in X_i DGPs with $X_i \sim U[-1, 1]$ where the threshold value is $c = 0$. For the features, I use four variables, two binary ($Z_{i,1-2}$) with 0.5 probability of being 1. For DGP-2 I add two uniformly distributed continuous variables: $Z_{i,3-4} \sim U[-5, 5]$.

DGP 1: Two treatment effect and homogeneous $\eta(\cdot)$. $Z_i = [Z_{i,1}, Z_{i,2}]$, where $Z_{i,1}$ is relevant for CATE, the other is irrelevant.

$$\eta(X_i) = 2 \times X_i$$
$$\kappa(Z_{i,1}) = Z_{1,i} - (1 - Z_{1,i})$$

DGP 2: Continuous treatment effect and heterogeneous $\eta(\cdot)$. $Z_i = [Z_{i,1}, Z_{i,2}, Z_{i,3}, Z_{i,4}]$, $Z_{i,3}$ is relevant for CATE, $Z_{i,2}$ has an effect on $\eta(\cdot)$, the others are irrelevant.

$$\eta(X_i, Z_{i,2}) = 2 \times Z_{i,2} \times X_i - 2 \times (1 - Z_{i,2}) \times X_i$$
$$\kappa(Z_{i,3}) = 2 \times Z_{i,3}$$

[45] I allow the splitting value for each feature to be within 0.5 thresholds to accept the split to be similar to the DGP's. Also, note that growing smaller or larger trees has different types of errors.

DGP 3-5 uses nonlinear specification for X_i. I follow Calonico et al. (2014) Monte Carlo setups, where $\eta(\cdot)$ is nonlinear in X_i and supplement with heterogeneous treatment effects. Calonico et al. (2014) imitate two empirical applications and add one extra setup to investigate the effect of excess curvature. For all three designs the running variable is generated by $X_i \sim (2\mathcal{B}(2,4)-1)$, where $\mathcal{B}$ denotes a beta distribution and the disturbance term has the variance of $\sigma_\epsilon^2 = 0.05$. The threshold value is the same as in DGP-1 and 2.

DGP 3: Imitating Lee (2008) vote-shares. I assume two treatment effects and heterogeneous $\eta(\cdot)$. I use 52 dummy variables representing political parties and states. Political party dummy ($X_{i,1}$) is relevant and has an effect on both treatment and functional form. States are irrelevant. For $Z_{i,1} = 1$, I set the functional form as in Calonico et al. (2014) first setup.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 0.48 + 1.27X_i + 7.18X_i^2 + 20.21X_i^3 + 21.54X_i^4 + 7.33X_i^5, & \text{if } X_i < 0, Z_{i,1} = 1 \\ 0.48 + 2.35X_i + 8.18X_i^2 + 22.21X_i^3 + 24.14X_i^4 + 8.33X_i^5, & \text{if } X_i < 0, Z_{i,1} = 0 \\ 0.48 + 0.84X_i - 3.00X_i^2 + 7.99X_i^3 - 9.01X_i^4 + 3.56X_i^5, & \text{if } X_i \geq 0, Z_{i,1} = 1 \\ 0.48 + 1.21X_i - 2.90X_i^2 + 6.99X_i^3 - 10.01X_i^4 + 4.56X_i^5, & \text{if } X_i \geq 0, Z_{i,1} = 0 \end{cases}$$

$$\kappa(Z_{i,1}) = 0.02 \times Z_{1,i} + 0.07 \times (1 - Z_{1,i})$$

DGP 4: Ludwig and Miller (2007) studied the effect of Head Start funding to identify the program's effects on health and schooling. I assume continuous treatment effect based on the age of participants. Age is assumed to be uniformly distributed: $Z_{i,1} \sim U[5,9]$ and I add dummies representing different continents involved in the analysis.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 3.71 + 2.30X_i + 3.28X_i^2 + 1.45X_i^3 + 0.23X_i^4 + 0.03X_i^5, & \text{if } X_i < 0 \\ 3.71 + 18.49X_i - 54.81X_i^2 + 74.30X_i^3 - 45.02X_i^4 + 9.83X_i^5, & \text{if } X_i < 0 \end{cases}$$

$$\kappa(Z_{i,1}) = -5.45 - (Z_{1,i} - 5);$$

DGP 5: 'An Alternative DGP' by Calonico et al. (2014) adds extra curvature to the functional form. This design is exactly the same as in Calonico et al. (2014), thus it has a homogeneous treatment effect. The features are the same as in DGP 4 and they are all set to be irrelevant.

Treatment effect and $\eta(\cdot)$ is homogeneous.

$$\eta(X_i, Z_{i,1}) = \begin{cases} 0.48 + 1.27X_i - 0.5 \times 7.18X_i^2 + 0.7 \times 20.21X_i^3 \\ \quad +1.1 \times 21.54X_i^4 + 1.5 \times 7.33X_i^5, & \text{if } X_i < 0 \\ 0.48 + 0.84X_i - 0.1 \times 3.00X_i^2 - 0.3 \times 7.99X_i^3 \\ \quad -0.1 \times 9.01X_i^4 + 3.56X_i^5, & \text{if } X_i < 0 \end{cases}$$

$$\kappa = 0.04$$

F Monte Carlo simulation for fuzzy design

For fuzzy designs, I use the same functional forms and setups for the DGPs, but add a homogeneous first-stage for getting the treatment:

$$T_i = \begin{cases} \mathbb{1}\left(0.5 + 0.8X_i + \nu_i > 0\right) \ , & \text{if } X_i \geq 0 \\ 0 & \text{if } X_i < 0 \end{cases}$$

where $\nu_i \sim \mathcal{N}(0, 1)$. For simplicity I use 'DGP-x-f' expression for referring these fuzzy setups. Table 3 and 4 show the same algorithm performance measures and the evidence on valid inference similarly to the sharp design. The results are aligned with the conclusion reported in Section 4, but the fuzzy design is even more data intensive.

DGP	N	inf. MSE	$\#\hat{\Pi}$	DGP found (%)
DGP-f-1	$N = 1,000$	1.1129	1.00	0%
	$N = 5,000$	0.0267	2.04	96%
	$N = 10,000$	0.0126	2.03	97%
DGP-f-2	$N = 1,000$	13.1595	2.00	-
	$N = 5,000$	4.6662	5.83	-
	$N = 10,000$	3.3652	8.99	-
DGP-f-3	$N = 1,000$	0.0012	1.00	0%
	$N = 5,000$	0.0003	1.99	99%
	$N = 10,000$	0.0001	2.00	100%
DGP-f-4	$N = 1,000$	1.6566	1.00	-
	$N = 5,000$	0.2255	3.00	-
	$N = 10,000$	0.1351	3.69	-
DGP-f-5	$N = 1,000$	0.0006	1.00	100%
	$N = 5,000$	0.0001	1.03	97%
	$N = 10,000$	0.0001	1.02	98%

Table 3: Monte Carlo averages for performance measures in fuzzy designs

Number of true leaves: $\#\Pi_{DGP-1} = 2$, $\#\Pi_{DGP-3} = 2$, $\#\Pi_{DGP-5} = 1$

Algorithm setup: using the smallest cross-validation value to select γ^*,

$q = 1$ for DGP 1 and 2 and $q = 5$ for DGP 3,4 and 5.

DGP 1	Leaf	$\ell_1 : \tau_1(Z_1 = 1) = 1$		$\ell_2 : \tau_1(Z_1 = 0) = -1$	
	Estimates	average bias	actual 95% CI coverage	average bias	actual 95% CI coverage
	$N = 1,000$	-	-	-	-
	$N = 5,000$	-0.0147	0.95	-0.0037	0.95
	$N = 10,000$	-0.0038	0.95	0.0020	0.96
DGP 3	Leaf	$\ell_1 : \tau_1(Z_1 = 0) = 0.07$		$\ell_2 : \tau_1(Z_1 = 1) = 0.02$	
	Estimates	average bias	actual 95% CI coverage	average bias	actual 95% CI coverage
	$N = 1,000$	-	-	-	-
	$N = 5,000$	-0.0002	0.96	0.0004	0.96
	$N = 10,000$	-0.0003	0.95	-0.0003	0.94
DGP 5	Leaf	Homogeneous Treatment, $\tau = 0.04$			
	Estimates	avgerage bias		actual 95% CI coverage	
	$N = 1,000$	-0.0000		0.95	
	$N = 5,000$	0.0001		0.96	
	$N = 10,000$	-0.0003		0.95	

Table 4: Estimated Monte Carlo average for bias and actual 95% confidence intervals coverage for each leaf for tree structured DGPs, conditional on DGP is found - fuzzy design

Note: For DGP-f-1 and DGP-f-3, with $N = 1,000$, there is no case when the true DGP is found, thus no values are reported.

G Additional results on the empirical exercise

This part adds additional information to the empirical analysis. Table 5 shows the descriptives for the used variables in the heterogeneity analysis. Here, I only present the variables used for revisiting the heterogeneity analysis by Pop-Eleches and Urquiola (2013).

	School level average transition score	Baccalaureate taken	Baccalaureate grade	Scaled School admission score	number of schools in town
Mean	7.65	0.74	8.12	0.10	17.50
Median	7.55	1.00	8.15	0.15	17.00
Std deviation	0.75	0.44	0.90	0.55	7.49
Min	5.78	0.00	5.19	-1.00	2.00
Max	9.63	1.00	10.00	1.00	29.00
N	1,857,376	1,857,376	1,256,038	1,857,376	1,857,376

Table 5: Descriptive statistics of the variables used in heterogeneity analysis of Pop-Eleches and Urquiola (2013)

Table 6 restates the main findings of Pop-Eleches and Urquiola (2013) on the heterogeneity exercise.

	School level average transition score	Baccalaureate taken	Baccalaureate grade
Full sample			
τ_0	0.107***	0.000	0.018***
$SE(\tau_0)$	(0.001)	(0.001)	(0.002)
N	1,857,376	1,857,376	1,256,038
Top tercile			
τ_1	0.158***	0.003	0.048***
$SE(\tau_1)$	(0.002)	(0.002)	(0.003)
N_1	756,141	756,141	579,566
Bottom tercile			
τ_2	0.099***	−0.008*	−0.005
$SE(\tau_2)$	(0.003)	(0.004)	(0.009)
N_2	392,475	392,475	212,282
Towns with four or more schools			
τ_1	0.097***	0.000	0.016***
$SE(\tau_1)$	(0.001)	(0.001)	(0.002)
N_1	1,806,411	1,806,411	1,223,341
Towns with three schools			
τ_2	0.333***	−0.007	0.028*
$SE(\tau_2)$	(0.007)	(0.009)	(0.016)
N_2	31,149	31,149	19,877
Towns with two schools			
τ_3	0.697***	0.020	0.179***
$SE(\tau_3)$	(0.010)	(0.013)	(0.023)
N_3	19,816	19,816	12,820

Notes: All regressions are clustered at the student level and include cutoff fixed effects. Standard errors are in parentheses. All estimates present reduced form specifications where the key independent variable is a dummy for whether a student's transition score is greater than or equal to the cutoff.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table 6: Heterogeneity in Baccalaureate Effects - Pop-Eleches and Urquiola (2013), Table 5

Table 7 summarize the different treatment effects estimated by Pop-Eleches and Urquiola (2013) and by the algorithm for Baccalaureate exam grade. Note that for RD tree: only number of schools, I only used the number of schools only as features. RD tree: all variables are using both average

transition score for the class and number of schools as features, but finds only average transition score variable as relevant.

	Avg. transition score for the class		Number of schools		
Pop-Eleches and Urquiola (2013)	Top tercile 0.048***	Bottom tercile −0.005	2 0.179***	3 0.028*	4-27 0.016***
RD tree: all variables	Below median† 0.015**	Above median† 0.028***	-		
RD tree: only no. schools	-		2 0.152***	3-24 and 26-27 0.021***	25 −0.013

Regressions are clustered at the student level and include cutoff FE.

***: significant at 1%, **: significant at 5%, *: significant at 10%.

†: the algorithm splits at 44th percentile.

Table 7: Heterogeneity in treatment effects for Baccalaureate grade

	Mean	Median	Std. dev.	Min	Max	N
Outcome and running variables						
School level average transition score	8.20	8.29	0.60	6.53	9.41	11,931
Scaled Admission score	0.85	0.82	0.97	-2.07	3.91	11,931
Socioeconomic characteristics of households						
Female head of household (d)	0.89	1	0.32	0	1	11,931
Age of head of household	46.75	45	7.15	13	97	11,843
Romanian (d)	0.94	1	0.24	0	1	11,931
Hungarian (d)	0.05	0	0.22	0	1	11,931
Gypsy (d)	0.01	0	0.06	0	1	11,931
Other Ethnicity (d)	0.01	0	0.09	0	1	11,931
HH's Primary education (d)	0.66	1	0.47	0	1	11,840
HH's Secondary education (d)	0.20	0	0.40	0	1	11,840
HH's Tertiary education (d)	0.13	0	0.34	0	1	11,840
Socioeconomic characteristics of students						
Gender of student (d)	0.42	1	0.49	0	1	11,931
Age of student	18.08	18	0.94	14	23	11,866
Accessibility of households to goods						
Car (d)	0.57	1	0.49	0	1	11,820
Internet (d)	0.73	1	0.44	0	1	11,829
Phone (d)	0.47	0	0.50	0	1	11,807
Computer (d)	0.87	1	0.34	0	1	11,851
Parental and Child responses to survey questions						
Parent volunteered (d)	0.11	0	0.31	0	1	11,868
Parent paid tutoring (d)	0.24	0	0.42	0	1	11,931
Parent helps HW (d)	0.20	0	0.40	0	1	11,815
Child does HW every day - Parent (d)	0.75	1	0.43	0	1	11,779
Negative interactions with peers	0.12	0	0.37	0	5	11,838
Child does HW every day - Child (d)	0.63	1	0.48	0	1	11,908
HW percieved easy	5.45	5.60	1.02	1	7	9,628
Characteristics of schools						
No. schools	2.33	2	0.50	2	4	11,931
2 schools (d)	0.69	1	0.46	0	1	11,931
3 schools (d)	0.29	0	0.45	0	1	11,931
4 schools (d)	0.02	0	0.13	0	1	11,931
Highest certification teacher in school (d)	0.61	1	0.49	0	1	11,169
Novice teacher in school (d)	0.06	0	0.24	0	1	11,169

(d) indicates it is a dummy variable. 'HH' stands for household, 'HW' for homework.

Table 8: Descriptive statistics of the used variables for exploring heterogeneity in a survey-based dataset

Table 8 shows the descriptives for the candidate features used to find the tree shown by Figure 8.