Applying Generalized Random Forest to Analyzing the Heterogeneous Effects in Regression Discontinuity Design

Master's Thesis presented to the
Department of Economics at the
Rheinische Friedrich-Wilhelms-Universität Bonn

In Partial Fulfillment of the Requirements for the Degree of
Master of Science (M.Sc.)

Supervisor: JProf. Dr. Claudia Noack

Submitted in August 2024 by:
Yu-Hsin Chen
Matriculation Number: 3461265

Contents

List of Tables

1 Introduction

In recent years, advancements in technology have brought substantial growth in the amount of data collected and stored daily. On one hand, the unprecedented amount of data provides abundant information for researchers to explore. On the other hand, it also highlighted the limitation of some typical statistical models, particularly in the face of the problem of the ”curse of dimensionality.” In this case, the Random Forest algorithm, from Breiman (2001), shows its capability to deal with large amounts of variables and data points. People can easily apply regression forests or classification forests for the purpose of prediction or classification. Moreover, in terms of treatment effects estimation, based on the theoretical works on the Random Forest, Wager and Athey (2018) and Athey et al. (2019) proposed a generalized Random Forest framework, which can incorporate several other treatment effects estimation methods, like instrument variables or quantile regression, to makes the treatment effects estimation heterogeneous. The aim of the thesis is to build a variety of regression discontinuity models on this heterogeneous effects estimation framework theoretically and applicably and, ultimately, bridge the gap between theory and application. This thesis can be divided into three main parts: theory, Monte Carlo simulation, and application. In terms of the theoretical section, first, some essential assumptions of regression discontinuity design for the generalized random forest framework will be introduced. It is mainly based on the research in Imbens and Lemieux (2008), also from Cattaneo et al. (2017). First, a brief introduction will be given in the theory section of the Generalized Random Forest(GRF). Other than that, because the thesis is based on the local regression model assumptions, some of the original assumptions from Athey et al. (2019) may not be necessary or can be less restrictive. Therefore, these assumptions will be rewritten and compared with the original assumptions from Athey et al. (2019). Furthermore, some proofs regarding asymptotic properties can be simplified; this will be done in the thesis. In the process, results from Wager and Athey (2018) and MeinshausenNicolai (2006) would be applied. To provide a detailed introduction to generalized random forest(GRF), the structure of a generalized random forest can be roughly separated into two sections: partitioning and estimation. For estimation, it is different from a typical random forest. Since the solutions for typical random forest to eliminate bias by averaging the result from each tree will not alleviate the bias for GRF. Therefore, in GRF, we adopt the forest as a weighting method. Those observations that fall into the same leaf as our interest covariates would obtain weights. In the end, we applied the local weighted regression in the moment function to estimate our interest parameter(s). The idea of adaptive locally weighted estimations is largely contributed by Hothorn et al. (2002) and Mein-

shausenNicolai (2006). Additionally, for partitioning, GRF does not apply MSE, Gini index, or entropy as a classical random forest. Differently, GRF has special partitioning criteria trying to maximize the heterogeneous effects between nodes and, at the same time, encourage the balanced allocation of individuals. With this setup, each leaf becomes a quasi-randomization experiment, and all the observations have similar characteristics (covariates). In other words, the difference of individuals in the same leaf shall ideally only be based on whether they receive the treatment effects or not. Consequently, we not only achieve randomization but, at the same time, reach the goal of heterogeneity from the difference of characteristics between leaf and leaf. The partitioning method in GRF is based on the research from Athey and Imbens (2016) and (2009) Su et al. (2009), which is mainly targeting the spitting criteria for treatment effects estimation. Additionally, a gradient-based method would be applied in the partitioning process to lower the computational cost.

Next, the Monte-Carlo simulations section can be divided into two components. First is the classical Monte-Carlo simulation, which aims to test how the different levels of complexity of the data-generating process affect different models in terms of sample size, noise, and irrelevant variables. On top of that, detailed simulation and analysis will be conducted based on different points along the covariates. The second will be based on Wasserstein generative adversarial neural networks(WGAN). The research is mainly based on Goodfellow et al. (2014), Gulrajani et al. (2017), and Athey et al. (2024). Originally, Goodfellow et al. (2014) introduced the concept of adversarial neural networks. The term implies that there are two neural networks in the process. One is the discriminator, the other is the generator, and the objective of the discriminator is to determine whether the data is generated by the generator. The goal of the generator is exactly the opposite: trying to deceive the discriminator as much as possible. Due to the instability issue, Gulrajani et al. (2017) proposed the idea of applying Wasserstein distance for estimation. Athey et al. (2024) combine the concept of WGAN with Monte-Carlo simulation. In this section, the real-world dataset will be applied as input to generate the estimated data, and only the estimated data will be used in the Monte Carlo simulation. A clear theoretical introduction about WGAN will be given; however, for the purpose of cohesion and coherence of the thesis, it will be included in the appendix; please reference the section WGAN. In the thesis, the dataset from Ludwig and Miller (2007a) would be applied for the second stage of the Monte-Carlo simulation, which is about the treatment effects of the Headstart program. The discussion in Ludwig and Miller (2007a) is about whether support from the Headstart program can decrease the mortality rate of children. Additionally, we will have a further extension in a more general heterogeneous effect estimation case without a specific running variable. The

dataset from Connors et al. (1996) would be applied. The paper's topic is about the effects of right heart catheterization on mortality. But to limit the discussion of the thesis specifically on the RDD context in the main part, the extension of the Monte-Carlo simulation will only be included in the appendix; please reference Extension: Second Stage Monte Carlo Simulation in a general case. The objective in the second stage of Monte Carlo simulation is first to apply the relatively new method (Monte Carlo Simulation on WGAN) and observe its result in mimicking very different data styles. At the same time, we will have a certain level of understanding of the original data, which will benefit the real-world data analysis in the last section. Moreover, the model's performance from the estimated real-world dataset will be observed, and different parameter settings will be tested. In the application part, the true datasets will be applied again. The GRF estimation result of treatment effects in the Headstart program will be compared with the result in the paper Cattaneo et al. (2017), which applied the typical regression discontinuity design methods. In the appendix, we can also find the extension application for GRF estimation in a more general context. The data from a dataset from Connors et al. (1996) would be applied again. The analysis result will be compared with the result from McConnell and Lindner (2019), and the heterogeneous effects analysis will be conducted.

2 Regression Discontinuity Design Assumptions and Functions

2.1 Parametric Assumption

This thesis will adopt the parametric assumptions for the models within the generalized random forest in the theory section. Therefore, the considerations of optimal bandwidth selection and kernel selection would not be included in the thesis. By not taking the bandwidth, the assumptions can fully coincide with the foundational assumptions present in the paper Athey et al. (2019). In contrast, specifically limiting the scope to observations within the chosen bandwidth may affect the convergence rate of our interest variables in the generalized random forest toward the true value because the parameters that determine the convergence rate in the formula are about sample size, sub-sample size, fraction of observations from parent nodes to child. Furthermore, optimal bandwidth selection is a relatively complex topic, and simply applying the MSE optimal bandwidth may lead to a biased result that may not be able to achieve standard normality theoretically. Therefore, in the theory section, this thesis would maintain the parametric assumptions for the model inside the GRF to secure consistency. Only in the ap-

plication part to compare to the result from Cattaneo et al. (2017), the Imbens-Kalyanaraman MSE-optimal bandwidth would be applied.

2.2 Regression Discontinuity Design Assumptions

Before delving into the specifics of the regression discontinuity design assumptions, let us clarify the notations used. First, X_i is the running variable, and Y_i is the outcome variable. D_i is a treatment variable that is binary.

Assumption 1:Continuity of potential outcome on running variables(X) conditionally on covariate(p):

When given the interest covariates "p", for all x in X will only makes y exhibit a jump at the cutoff point, which helps us to simplify the problem that we are focusing on estimating the treatment effect locating at cutoff point. There is no other jump to make the coefficient of the treatment variable biased. Additionally, note that in a generalized random forest, we will use the entire dataset to estimate the heterogeneous treatment effect. Therefore, the assumption here can be stronger than the assumption in typical RD design. For the classical regression discontinuity design, the continuity assumption of outcome variables on running variables may only hold around the cutoff point because the bandwidth filters many data points.

Assumption 2: Continuity of distribution of covariates in all running variable x belongs to X:

From the regression discontinuity design(RDD) perspective, this assumption can prevent the violation of the local randomization experiment. In RDD, we assume that the distribution of covariates near the cutoff point should not exhibit any jump. Not only should the expectation not exhibit a jump, but the variance also should not exhibit a jump. In the end, it implies that near the cutoff point, all the observations have similar characteristics, and their only difference is whether they receive the treatment effect or not. From the GRF perspective, we would use the covariates to split the tree. If there is a jump, it may lead to the case that a child node may contain either no treated observations or all treated observations, which is opposite from the objective of GRF. In GRF, we expect that in each leaf, all the observations have similar characteristics(covariates), like inside the bandwidth of typical RDD, and some of them receive the treatment, but others do not. So that, on one hand the treatment effect can be estimated, on the other hand, the heterogeneous effects are shown between leaf.

Assumption 3:There is no manipulation problem on running variable X:

This assumption is to prevent some cases in which the individuals know the threshold of

receiving/avoiding the treatment and take action before the treatment is conducted. This concept is presented by McCrary (2008). Without the assumption, it may lead to the case that some individuals are wrongly included or excluded from the bandwidth, affecting the treatment effect estimation. In the GRF, those observations may also cause the estimation problem in their leaf.

2.3 Regression Discontinuity Design Model

In our parametric model, four models will be applied. Two of them are linear, and the other two are polynomials. The exact forms are as follows.

$$Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \tag{1}$$

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i \tag{2}$$

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 X_i^2 + \varepsilon_i \tag{3}$$

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i D_i + \beta_5 X_i^2 D_i + \varepsilon_i \tag{4}$$

Equation one would only be applied in the extension discussion of heterogeneous effects estimation on general cases in Monte-Carlo simulation and application. Additionally, because equation(2) to equation(4) would be applied several times in the thesis, therefore in the rest of the paper, the equation(2) would be called as linear model, and equation(3) would be called as polynomial model, and equation(4) fourth would be called as PMWIT(polynomial model with interaction term).

3 Generalized Random Forest

In this section, the models from the previous paragraph will be integrated into the GRF framework. First, the details of combining regression discontinuity design with GRF will be introduced. Subsequently, the essential assumptions for the asymptotic properties of GRF will be discussed. This is because after knowing the mechanism of GRF, it will be easier to understand the necessity of the existence of those assumptions.

Note that we would keep the same notation of the running variables, outcome variable, and covariates as $(\mathrm{X}_i, Y_i, P_i) \in \mathscr{X} \times \mathbb{R}$ for this thesis. Next, the function would be written in

vector form; I further define them as follows. $Z_i = \begin{bmatrix} 1 \\ D_i \\ X_i \\ \vdots \end{bmatrix}$, the true Z_i would depends on which model is applied. For β again would be defined as $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \end{bmatrix}$. Subsequently, according to these notations, the main assumption of GRF, the first-moment condition of our regression discontinuity design model, also referred to as the score function in the paper Athey et al. (2019), would be established as follows.

$$\mathbb{E}\left[\psi_{\beta_1(p),\nu(p)}(Y_i) \mid P_i = p\right] = 0 \text{ for all } p \in \mathscr{P} \tag{5}$$

$$\psi = Z_i[Y_i - Z_i'\beta(p)] \tag{6}$$

Here, the true form of ψ would based on our linear model, polynomial model, and polynomial model with interaction term(PMWIT) in the vector form.

$$(\hat{\beta}_1(p), \hat{\nu}(p)) \in \arg\min_{\beta_1,\nu} \left\{ \left\| \sum_{i=1}^{n} \alpha_i(p)\psi_{\beta_1,\nu}(Y_i) \right\|_2 \right\} \tag{7}$$

Here β_1, the coefficient of D_i, is our parameter of interest; other beta would be called ν. α is the weight of the observations; details of it will be explained in the next paragraph. If there is a unit root for our main optimization problem (7), we can derive the result in the following way.

$$\sum_{i=1}^{n} \alpha_i(p) Z_i[Y_i - Z_i'\hat{\beta}(p)] = 0$$

$$\sum_{i=1}^{n} \alpha_i(p) Z_i Y_i = \sum_{i=1}^{n} \alpha_i(p) Z_i Z_i' \hat{\beta}(p)$$

$$\left[\sum_{i=1}^{n} \alpha_i(p) Z_i Z_i'\right]^{-1} \left[\sum_{i=1}^{n} \alpha_i(p) Z_i Y_i\right] = \hat{\beta}(p)$$

Also, we can centralize the result in the following form.

$$\left[\sum_{i=1}^{n} \alpha_i(p)(Z_i - \bar{Z_\alpha})(Z_i - \bar{Z_\alpha})'\right]^{-1} \left[\sum_{i=1}^{n} \alpha_i(p)(Z_i - \bar{Z_\alpha})(Y_i - \bar{Y_\alpha})\right] = \hat{\beta}(p)$$

$$as\ \bar{Z_\alpha} = \sum_{i=1}^{n} \alpha_i Z_i\ and\ \bar{Y_\alpha} = \sum_{i=1}^{n} \alpha_i Y_i$$

3.1 The weight of forest

The weight α is actually the **average** of α from each tree. First, we specify a **p** that we are interested in, and for each tree, all the observations which falls into the same leaf as p will obtain weight as the following form $\alpha_{b_i}(p) = \frac{1\{P_i \in L_b(p)\}}{|L_b(p)|}$. After we know each tree's weight, we have to re-average all of them to make the sum equal to one in a forest. $\alpha_i(p) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{b_i}(p)$

3.2 Tree growing process

All the trees still apply some common rules, like a normal random forest. During the splitting process, the greedy method would be applied. Also, only a subset of the sample and a subset of auxiliary covariates would be considered for each splitting round. The difference can be that we will apply the honest tree method. In short, the sub-sample would be further split to halve again; one halve would only be applied for splitting, and the other halve would be applied for weighting. The details of the honest tree can be found in Wager and Athey (2018). Regarding the splitting criteria, the GRF algorithm also differs from the typical Random Forest. Regarding classification trees, people usually apply the Gini index or entropy as a feature to quantify the purity of nodes and make further decisions about how to split. For regression trees, the focus is usually on identifying the feature and threshold that can minimize the MSE. However, because our results are based on the first-moment condition equation. Therefore, simply using those splitting criteria may have a chance of causing biased results. In order to find an alternative, at first, Athey et al. (2019) defined the splitting criteria as follows, trying to minimize the error function.

$$\text{err}(C_1, C_2) = \sum_{j=1,2} \mathbb{P}[P \in C_j | P \in Parent] \times \mathbb{E}\left[(\hat{\beta}_{1c_j}(J) - \beta_1(P))^2 \mid P \in C_j\right] \tag{8}$$

J means a giving sample. C_j means the training sample falls into the child node, $\beta_1(P)$ is the true parameter in child node C_j. And $\hat{\beta}_{1c_j}(J)$ is trained by child data. The parent stands for the observation in the parent node. Therefore, we can see that the error equation is, in fact, taking the average of the L2 distance between the true parameter and the estimated parameter in the children node. From Athey et al. (2019) proposition 1, they prove that this error function can be rewritten in the following form.

$$\text{err}(C_1, C_2) = K(parent) - \mathbb{E}[\Delta(C_1, C_2)] + o(r^2)$$

$$\Delta(C_1, C_2) := \frac{n_{c_1} n_{c_2}}{n_p} \left(\hat{\beta}_{1c_1}(J) - \hat{\beta}_{1c_2}(J) \right)^2 \tag{9}$$

As you can see, K(parent) is a function about the sample in the parent node, which has nothing to do with C_j. It is actually a variable related to $\text{Var}[\beta_1(P)]$ for P in the parent node, so here, it can be regarded as a deterministic term. Therefore, the term we need to analyze is $\Delta(C_1, C_2)$. Our goal turns to to maximize the heterogeneous effects, and we can also consider the problem as maximizing the parameter difference between two children nodes. On the other hand, it also encourages two nodes to contend with the same or similar sample size to maximize the fraction. However, calculating this optimization problem can be a very time-consuming task. Because to find the optimal split feature and threshold, we need to estimate several times β_1 and $\Delta(C_1, C_2)$, on top of that, we need to do it iteratively for each split, for each layer in the tree, and for all the trees in the forest.

Athey et al. (2019) suggest a gradient-based method to lower down the computational cost. First they use the following equation to approximate the $\Delta(C_1, C_2)$

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{|\{i : P_i \in C_j\}|} \left(\sum_{\{i: P_i \in C_j\}} \rho_i \right)^2$$

In Athey and Wagner(2019) Athey et al. (2019) proposition two states that it would tend to $\Delta(C_1, C_2)$ when sample size tends to infinity. Here ρ_i is from the connection between $\hat{\beta}_{1parent}$ and $\tilde{\beta}_{1c}$ like following form.

$$\tilde{\beta}_{1C} = \hat{\beta}_{1parent} - \frac{1}{|\{i : P_i \in C\}|} \sum_{\{i: P_i \in C\}} \xi^T A_{parent}^{-1} \psi_{\hat{\beta}_{1parent}, \hat{\nu}_{parent}(Y_i)}$$

$$\rho_i = \xi^T A_{parent}^{-1} \psi_{\hat{\beta}_{1parent}, \hat{\nu}_{parent}(Y_i)}$$

Here, we start from the end of the equation, and we can see that the parameters of ψ, our score function, which is $\hat{\beta}_{1parent}$ and $\hat{\nu}_{parent}$, both of them are from the parent node. We calculate them similarly to the (7) but without the weight α. So, the β would be like the following form.

$$\left[\sum_{i=1}^{n} (Z_i - \bar{Z}_{parent})(Z_i - \bar{Z}_{parent})' \right]^{-1} \left[\sum_{i=1}^{n} (Z_i - \bar{Z}_{parent})(Y_i - \bar{Y}_{parent}) \right] = \hat{\beta} \ \forall i \in \textit{Parent node}$$

Particularly because the parameter is based on the parent node, we do not need to calculate

them several times in the estimating process, which would save a lot of computational cost. Besides, A_{parent} is the gradient term of ψ, and it is composed of a gradient vector that points in the shortest direction to lower the score function. And the way to do it is by taking partial derivatives on β.

$$A_{parent} = \frac{1}{|\{i : P_i \in Parent\}|} \sum_{\{i:P_i \in parent\}} (Z_i - \bar{Z}_{parent})(Z_i - \bar{Z}_{parent})'$$

Additionally, ξ^T is a simple vector to pick the element in the matrix we are interested in, so it can be $\xi = \begin{pmatrix} 1 \\ 0 \\ \vdots \end{pmatrix}$ depends on our assumption. Until now, we have finished building up the generalized random forest. Based on the forest, we can calculate the weight for each observation and plug it into equation (7) and solve the optimization question there, then we will obtain our final estimation result. Next, we will discuss the asymptotic property of our main interest parameter $\hat{\beta}_1$

3.3 Assumption for asymptotic analysis

Assumption 4: Lipschitz Continuity of function $E[\psi_{\beta_1,\nu}(Y_i)|P = p]$ in p

Here, we assume that given a fixed value of β_1 and ν, the $E[\psi_{\beta_1,\nu}(Y_i)|P = p]$ function has to be Lipschitz continuous in covariates. To understand the assumption intuitively, we can say one reason is that we want to ensure that changing the covariates would not trigger the treatment effects. This can prevent the case that all the treated or control observations are allocated together in the splitting process. From the RDD point of perspective, we also expect that the covariates can not easily be distinguished near the border so that we are sure that the observations in the bandwidth have similar characteristics and their only difference is based on treatment effects. From the theoretical perspective, we need this assumption because many of the convergence properties of our interest variable are originally based on the change of p.

$$\mathbb{E}\left[\sup\{\|P_i - p\|_2 : \alpha_i(p) > 0\}\right] = O\left(s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log(\omega^{-1})}}\right) \tag{10}$$

This result allows us to derive the convergence rate of $\psi_{\beta_1,\nu}$. It is from Wager and Athey (2018). To provide a quick note, ω here is the least proportion of observations from the parent node inherited to the child node. π is the least proportion of features considered for splitting. Noting that here we cannot use it yet because there is another assumption that needs to be satisfied for applying the result, here is a quick view to see why it is essential to have Lipchitz

continuous for $\psi_{\beta_1(p),\nu(p)}$. Furthermore, because of our RD design models, we can easily satisfy some assumptions in the original paper. For example, from the original paper, they have the following assumption.

$$\gamma\big((\beta_1,\nu),(\beta_1',\nu')\big) \le L\left\|\begin{pmatrix}\beta_1\\ \nu\end{pmatrix}-\begin{pmatrix}\beta_1'\\ \nu'\end{pmatrix}\right\|_2 \quad \text{for all } (\beta_1,\nu),(\beta_1',\nu'),$$

$$\gamma\big((\beta_1,\nu),(\beta_1',\nu')\big) := \sup_{p\in\mathscr{P}}\left\|\mathrm{Var}\left[\psi_{\beta_1,\nu}(Y_i)-\psi_{\beta_1',\nu'}(Y_i)\mid P_i=p\right]\right\|_F$$

Here, they further restrict the sensitivity of the ψ function in terms of β_1 and ν, to make sure that even in the worse case of $p \in P$, the variance would still be a finite number in the Frobenius norm. But for our case, because our ψ function is Lipschitz continuous in β. Therefore it automatically holds.

Backing to assumption 4, given p, the prerequisite of assumption being held is to make sure the term in the β, as what we derived before, $E[X_i|P=p]$, $E[Y_i|P=p]$, $E[W_i|P=p]$, $\mathrm{Var}[Z_i|p]$ and $COV[Y_i,Z_i|p]$ are all Lipschitz continuous. Here, some of them exactly coincide with our previous assumptions. For example, the distribution of P should be continuous on X, and receiving the treatment or not is irrelevant to covariates,but note that the amount of treatment effects is largely depends on covariates so that the treatment effects can be heterogeneous.

Assumption 5: Assuming $\frac{\partial \psi_{\beta_1(p),\nu(p)}}{\partial\beta_1(p)\partial\nu(p)}$ is invertible

This assumption makes sure that the gradient of the ψ function not only exists but is invertible. This is crucial for our tree-splitting process. From the RDD model assumption we can already see that it is differentiable for the variable of interest β_1, also ν, which ensure that the gradient of ψ function exists. Now, we assume that the derivatives matrix that we got is not a singular matrix, so that we can obtain the term, A_{parent}^{-1} in the tree-splitting process.

Assumption 6: The existence of solution in main optimization problem(7)

In the original paper, Athey et al. (2019) regulate that the optimal interest parameter(s) from (7) need to at least satisfy the inequality

$$\left\|\sum_{i=1}^{n}\alpha_i\psi_{\hat{\beta}_1,\hat{\nu}}(Y_i)\right\|_2 \le C\ max\{\alpha_i\}$$

In the inequality, C is a constant. Note that in the ideal case, n tends to infinity, which implies that α_i would also tend to zero because countless observations would fall into the same leaf. Even C multiply the biggest α_i is still an extremely small value. Therefore, this assumption

implies that the left-hand side of the inequality should tend to zero as well.

Assumption 7: $\sum_{i=1}^{n} \alpha_i(p)\psi_{\hat{\beta}_1(p),\hat{\nu}(p)}$ is a negative gradient of a strong convex function:

This assumption is slightly different from the original paper. Athey et al. (2019) assume $\psi_{\hat{\beta}_1,\hat{\nu}}$ is the negative sub-gradient, in case that the ψ function is not differentiable at the some point. In contrast, for our case, it can be directly assumed that $\psi_{\hat{\beta}_1,\hat{\nu}}$ function is a negative gradient of a strong convex function. This assumption is a preparation for applying the property of strong convex function to prove that our estimated $\hat{\beta}_1$ would be closer to true β when the sample size increases. In the proving process, the fundamental theorem of calculus for line integrals will be applied, and the modified proof will be written in the appendix.

Assumption 8: The assumption of minimum subsample size

$$t_{\min} := 1 - \left(1 + \pi^{-1}(\log(\omega^{-1}))/\left(\log\left((1-\omega)^{-1}\right)\right)\right)^{-1} < t < 1$$

$$s = n^t \, for \; some \; t_{min} \leq t \leq 1$$

This assumption is based on the derivation result in MeinshausenNicolai (2006). A quick clarification of notation: ω means the minimum rate that the observations pass from the parent to the children. π means the minimum probability that the k-th feature is selected for splitting. s is the subsample size for building the random forest. This is in fact the prerequisite for using the result equation(10). In assumption 4, the continuity assumption of expectation of ψ in p, we were still not able to apply the result from Wager and Athey (2018). But now, after assumption 8 holds, we can use it for analyzing the asymptotic property of our ψ function.

3.4 Result from asymptotic analysis

Based on several assumptions above, we can now use the following theorem for generalized random forest. But before we delve into the theorem part, because it is hard to directly analyze the property of estimator,Athey et al. (2019) re-write our interest variable β in a another form.

$$\tilde{\beta_1}^*(p) := \beta_1(p) + \sum_{i=1}^{n} \alpha_i(p)\gamma^*(p)$$

$$\gamma^*(p) = -\xi V(p)^{-1}\psi_{\beta_1(p),\nu(p)}(Y_i)$$

Here is actually very similar to our tree splitting process, but notice that here $\tilde{\beta_1}^*$ is for the whole forest, also $\beta_1(p)$ is the true parameter of the observation p. Because of this linear assumption of our interest variable, we can use some results from Wager and Athey (2018) for asymptotic

property analysis.

Result 1: Based on the assumption 4-8, the estimated parameter $(\hat{\beta}_1, \hat{\nu})$ would converge in probability to true parameter (β_1, ν)

First, to further derive the convergence rate of the interest parameter(s) to the true parameter(s), we will start from the result (10) mentioned in Wager and Athey (2018) and in the deriving process, the property of Lipschitz continuous is essential. Furthermore, rely only on assumption 4 is not enough, we need the assumption 8 to be held to apply the result. Based on the property of Lipschitz continuity, we can derive the following convergence rate, and to have the weight, we need assumption 5 to be satisfied.

$$\left\| \mathbb{E}\left[\sum_{i=1}^{n} \alpha_i(p) \psi_{\beta_1(p),\nu(p)} \right] \right\|_2 = O\left(s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log(\omega^{-1})}} \right) \tag{11}$$

From the Wager and Athey (2018),(10), they state that that among all observations that their weight α_i bigger than zero, in the worst case of p, they still will have the convergence rate $s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log(\omega^{-1})}}$. Now, based on the Lipschitz continuous, we can expand this result to the whole ψ function. Next, based on the results of Hoeffding (1948), we can obtain the following result about the inequality variance of a single tree and forest.

$$\frac{n}{s}\mathrm{Var}\left[\sum_{i=1}^{n} \alpha_i(p) \psi_{\beta_1(p),\nu(p)} \right] \leq \mathrm{Var}\left[\sum_{i=1}^{n} \alpha_{bi}(p) \psi_{\beta_1(p),\nu(p)} \right] = O(1) \tag{12}$$

Here, we can see that on the right-hand side of the inequality is a single tree, and on the left-hand side is the forest; the variance of the forest will be smaller or equal to $\frac{s}{n}$ of variance in a tree. Based on this result, we can further derive the variance as follows:

$$\left\| \mathrm{Var}\left[\sum_{i=1}^{n} \alpha_i(p) \psi_{\beta_1(p),\nu(p)} \right] \right\|_F = O\left(\frac{s}{n}\right) \tag{13}$$

Based on the result of (11) and (13), we know that under the true parameter β_1 and ν, the function $\sum_{i=1}^{n} \alpha_i(p) \psi_{\beta_1(p),\nu(p)}$ will converge to zero. After this, we will discuss the case of our estimate variable in the function. $\sum_{i=1}^{n} \alpha_i(p) \psi_{\hat{\beta}_1(p),\hat{\nu}(p)}$. Now, we need assumption 6, the existence of a solution. From assumption 6 we know that function with estimated result will also converge to zero when the sample size tends to infinity. Intuitively, we can consider that because both functions would converge to zero, so the estimated function $\psi_{\hat{\beta}_1(p),\hat{\nu}(p)}$ may converge to $\psi_{\beta_1(p),\nu(p)}$. Then $\hat{\beta}_1(p), \hat{\nu}(p)$ would converge to $\beta_1(p), \nu(p)$. The formal proof is written in the appendix; please check Proof of Result 1. Besides, in the derivation process, assumption 7 will

be applied, therefore 4-8 assumptions are all necessary for result 1, and based on result 1 we can further derive result 2

$$\textbf{Result 2: } \sqrt{\frac{n}{s}}\left(\tilde{\beta_1}^*(p)-\hat{\beta_1}(p)\right)=O_P\left(\max\left\{s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log(\omega^{-1})}},\left(\frac{s}{n}\right)^{\frac{1}{6}}\right\}\right)$$

Because we plan to use $\tilde{\beta_1}^*$ to replace $\hat{\beta_1}$ in asymptotic properties analysis, it is particularly important to prove that they are very close to each other. The original proof is more complicated because Athey et al. (2019) considers some cases like quantile regression that their ψ function is not differentiable, leading to a situation where the gradient cannot be calculated. In the end, they cannot directly apply the Taylor expansion. However, for our case, the gradient for the ψ the function is estimable; therefore the technical proof is re-written and simplified and put in the appendix; please check Proof of Result 2

$$\textbf{Result 3: } \frac{\hat{\beta}_{1n}(p)-\beta_1(p)}{\sigma_n(p)}\sim\mathcal{N}(0,1)$$

First, because the form of $\tilde{\beta_1}^*$ is identical to the one in Wager and Athey (2018). And, they already proved that $\tilde{\beta_1}^*$ has Guassianity with σ_n. Besides, we have the rate how $\tilde{\beta_1}^*$ and $\hat{\beta_1}$ close to each other when n, s, ω or π changes. And according to that rate, compared with the lemma 7 in Wager and Athey (2018), we know that $\tilde{\beta_1}^*-\hat{\beta_1}$ would converge faster than the σ_n, which imply that $\tilde{\beta_1}^*$ and $\hat{\beta_1}$ is closer than $\tilde{\beta_1}^*$ and β_1. Moreover, we have the consistency of $\tilde{\beta_1}^*$. Therefore, it implies that it must exist an $\sigma_n(p)$ such that makes the result Guassianity hold. After knowing exist a Guassianity, we still need to find the $\sigma_n(p)$ for normality and confidence interval, but the process would be exactly the same as Athey et al. (2019), so it will be skipped in the thesis. In short, the process will be based on the methods including variance decomposition, delta method, and U-statistics calculating and again would rely on the coupling between $\hat{\beta_1}$ and $\tilde{\beta_1}^*$.

4 Monte Carlo Simulation

The Monte-Carlo simulation is comprised of two stages. The first stage aims to apply the finite simulated sample size to build up the connection between theory and application. In the first simulation stage, regression discontinuity design context will be mimicked while ensuring heterogeneous effects across observations according to their covariates. In this setting, we can observe the GRF capacity to estimate the HTE and also notice some potential issues of GRF. In the process, three sub-scenarios with different difficulty levels for estimating HTE will be

included. In each sub-scenario, the effects from the covariates will have a different way of entering the data-generating process. It would also change from linear to non-linear, and the running variables may also affect outcome variables directly and indirectly(through treatment). Besides, three regression discontinuity design models align with our model assumption: the linear model, polynomial model, and polynomial model with the interaction term would be tested. In the second stage of the Monte-Carlo simulation, Wasserstein generative adversarial networks will be applied. It would be mainly based on the paper, Athey et al. (2024), Goodfellow et al. (2014), and Gulrajani et al. (2017). Furthermore, because the GAN model applies real-world data as input, the data for the thesis will be from Ludwig and Miller (2007b) about the Headstart program. Besides that, we can find an extension for a more general HTE estimation case with more variables and complex context from the appendix; Extension: Second Stage Monte Carlo Simulation in a general case.The data from Connors et al. (1996) will be applied

4.1 First stage Monte-carlo simulation

In this section, the notation will be defined as follows. D is the binary treatment variable. X is running variables that follow the standard normal distribution, and $x = 0$ is the cutoff point. All the observations with $x > 0$ would receive the treatment. Additionally, there are two covariates P_1 and P_2, and both follow the standard normal distribution. Moreover, the error term also follows a standard normal distribution. Note that running variables, covariates, and error terms are all independent to each other, so the assumption should be held. Next, we have three data-generating processes were designed and applied. First is the simple case.

$$Y_i = D_i(P_{1i} + P_{2i}) + 2X_i + P_{1i} + P_{2i} + \varepsilon_i$$

From this straightforward data-generating process, we can discern several key details. First, it exhibits heterogeneous treatment effects influenced by P_{1i} and P_{2i} across the entire dataset. Additionally, during the tree-splitting process, GRF would apply the covariates P_{1i} and P_{2i}. Both of them are uncorrelated to the treatment D_i, so from the design of the splitting process, ideally, each leaf should contain a balanced number of treated and control observations. Also, individuals with similar characteristics should be allocated to the same leaf, making the leaf like a quasi-randomization environment so that the RD design model can estimate their treatment effects precisely. Other than that, the treatment effects should be different from leaf to leaf, which, in the end, can reach our goal, the heterogeneous treatment effects estimation. Additionally, the design of the data-generating process shows that the covariates not only impact the

outcome variable Y_i from treatment, but they also directly affect the outcome variables Y_i. Also the running variables has a direct impact on Y_i, instead of just from treatment effect decision, these may also increase the complexity to estimates the treatment effect precisely.

Next, we have the medium-level data-generating process.

$$Y_i = 2 + D_i(1 + P_{1i} + P_{2i} + P_{1i}^2 + P_{2i}^2) + 2X_i + 1.5X_i^2 + P_{1i} + P_{2i} + \varepsilon_i$$

From simple to medium level, P_{1i}^2 and P_{2i}^2 terms are added into the data generating process, and both of them follow standard normal distribution. Therefore, the splitting process now has to take into account a more heterogeneous case. Because the treatment effects would also be affected by the squared term, therefore, to estimate precisely, it may need some extra layers to deal with this case. Furthermore, the X_i^2 added in the data-generating process may also make the coefficient estimate of X_i harder and indirectly affect the treatment effects estimation, especially in the case when the RD model is not specified well. For the complex sub-scenario, we have the following DGP.

$$Y_i = 2 + D_i(1 + P_{1i} + P_{2i} + P_{1i}^2 + P_{2i}^2 + X_i + X_i^2) + 2X_i + 1.5X_i^2 + P_{1i} + P_{2i} + \varepsilon_i$$

Increasing the difficulty from medium to complex, the X_i and X_i^2 are incorporated into the treatment effects. These terms would not make the treatment effects more heterogeneous because X is independent with covariates, but the incorporation complicates the estimation process within each leaf. Without specifying D_iX_i and $D_iX_i^2$ correctly in the RD design model, it may reflect on the mean squared error, bias, standard error, and confidence interval.

Next, the sample size, noise, and number of irrelevant covariates would be tested in each sub-scenario(simple, medium, and complex) simulation. There are four different sample sizes, 1000, 2000, 3000, and 4000, that will be tested. For noise, it is actually the variance of noise terms that follow the normal distribution. It will start from 1, as a benchmark, and then 1.5,2,2.5 respectively. For irrelevant covariates, those covariates will be mixed with the real covariates P_1 and P_2 and taken into account in the splitting process, but in fact, they do neither contribute to the treatment nor outcome variable, and they also follow the normal distribution. The number of noise variables will start from 0, which we can take as a benchmark, and then gradually grow to 5, 15, and 25, respectively. The estimation criteria would be root mean squared error, standard error and 95% confidence interval coverage.

Furthermore, the first stage simulations will be further divided into two sections for noise

variables test and sample size test. In the first section, there are 1000 observations will be drawn from the data-generating process. In the iteration and estimation, these 1000 observations will be fixed from beginning to end, and for each iteration, new data will be drawn from the data-generating process only for training purposes so that the simulation result can coincide with the result of Gaussian normality and confidence interval coverage that we derive in result 3. In the further section, we will closely examine the performance of different points along the covariates, and because both covariates follow a standard normal distribution, the testing point will be set as -1, -0.5, 0, 0.5, 1. This approach allows us to clearly observe the performance of GRF at varying points on covariates. The RMSE, mean, standard error, and confidence interval will be calculated for evaluation purposes. For easy comparison reasons, long tables and similar tables in the simulation will be put in the appendix. Last but not least, in the experiment, all the general cases are done in the default settings of generalized random forest parameters not to benefit any points or context in the experiment. It is done with 2000 trees, at least 5 observations in the leaf, the subsampling rate is 0.5, and the honest fraction is also 0.5, and will consider $\sqrt{num(p)}+20$ variables each time.

From the beginning, we start with the straightforward test, which is the result of the noise experiment. In the appendix, we can find the corresponding table titled "Noise Experiment Result In The General Case". The benchmark of all the tables has the following setup: noise = 1, sample sizes = 1000, and number of irrelevant covariates = 0 for all the general tables. When we delve into the table, from the RMSE part, we can see a discernible increasing trend overall. From the simple sub-scenario, we see that the linear model and polynomial model have similar results, which implies that when there is an additional X^2 added in the model, for this DGP, it did not cause much effect on estimation. However, when introducing two additional redundant terms, D_iX_i and $D_iX_i^2$, may cause a slightly stronger impact on the treatment effect estimation. This exhibit in the polynomial model with interaction term(PMWIT) section has a higher RMSE than the other two in a simple case. For the mean standard error, we can observe an obvious increasing trend. It is result from that the variability of noise terms makes the estimation unstable between tree and tree. Two additional terms from PMWIT models may also make it more unstable than others. For the confidence interval coverage, there is a clear increasing trend for most of the cases. Because changing the noise parameter may not directly cause significant biased results, but it makes the estimation unstable which increases the standard error. In the end, the coverage interval becomes wider, which then leads to a higher coverage rate. Furthermore, because the linear model and polynomial model have more precise specifications, therefore they have a higher coverage rate than PMWIT. Additionally, here we

may doubt why, even in the simple data-generating process with reasonable model specification, the coverage rate for all the models is still only around 90%. This point will be further discussed in the irrelevant variables experiment, and we will have a clearer view after observing the result along different points on our covariates P_1 and P_2.

In the medium scenario, we can see that the linear model's absence of the $X2$ term results in a larger RMSE than other models. Additionally, the polynomial model's performance is slightly better than PMWIT's in terms of RMSE and confidence coverage rate due to its well-specified structure.

In the complex scenario, we can again see that PMWIT has a more precise outcome than the other two models when it is benchmark noise 1 because it specified the D_iX_i and $D_iX_i^2$ term in the RD design model. And for the polynomial model, at least it specified the X^2, so its RMSE is still lower than a linear model. For mean, standard error, the polynomial model shows a relatively low value, which is even lower than PMWIT. The reason can be that it underestimates the standard error because the X and X^2 terms absorb the treatment effect, which may create a fake consistency between trees. It also directly reflects on the 95% coverage rate. It only has around 75 % coverage rate in the beginning. In contrast, linear model and PMWIT had a better performance for the coverage rate. At the end of this table, we can learn that the misspecification may not necessarily reflect on RMSE directly, but it also can reflect on the underestimation of mean standard error and lead to a relatively poor confidence interval coverage rate.

Next, please reference the Noise Variables Experiment Result In The General Case in the appendix. Our objective in this section is to observe the effects caused by redundant covariates for different model settings. These covariates do not contribute to outcome variable Y or the treatment effect. We start with a simple sub-scenario and linear model. The linear model is correctly specified for the simple case; as the number of noise covariates increases, there is a clear, discernible upward trend for RMSE. However, we need to carefully interpret the result here that it does not imply when the number of noise variables increases, all the points on covariates will show an increasing trend. In comparison, the result here is a kind of "average result" among all the points from the data-generating process. Later, we will see how the noise variables affect different points of covariates. For the polynomial model, it shows almost exactly the same pattern as the noise test table shown before. For the polynomial model with an interaction term(PMWIT), again, the same as the previous experiment, it has a slightly higher RMSE than the other two models in the simple case. In the coverage rate, we can again see the trend that an increase in the number of irrelevant covariates may decrease the coverage rate "in general."

For the medium case, the linear model does not exhibit a straightforward increase trend in RMSE. Instead, it climbs to 1.745 and drops to 1.743, even though there are ten more noise variables added. In the polynomial models, it shows a similar pattern; from 0 noise variables to 5 noise variables, the RMSE shows an increasing trend, and then it goes down to 1.221. In a PMWIT, as the number of noise variables increases from 15 to 25 noise variables, it again does not show a significant increase. Furthermore, in the complex case, the RMSE of linear model shows a drop when from 0 noise variables to 5 noise variables. Based on these results, we can doubt that noise variables may not just cause detrimental effects for all the points in this context, but only in general, it leads to an increasing RMSE trend. Also, according to the setup of the data-generating process, we know that both covariates follow standard normal distributions, which means that the majority of data points are located around the center. In the experiment process, I do not specifically change the parameters of GRF for any point. Therefore, we can reasonably suspect a substantial over-fitting issue at the center, but it gradually gets less when we move from the center to the boundary. To understand it intuitively, we can say that because a huge number of data points are in the middle of the tree-splitting process, they will be split several rounds until only a few are left in the leaf. However, for those observations closer to the boundary point, like -1 and 1(around 15% and 85% quantile), because there are significantly fewer than the majority, after splitting once or twice, there are already very few observations left in the leaf, then it has to stop splitting further. Therefore, there are fewer or no over-fitting problems for them. In this case, increasing the noise variables may cause a huge detrimental impact on them regarding RMSE and confidence interval coverage.

To further prove this guess, a table with various points on the covariates has been made, and it is based on a simple scenario DGP. Noting that, to estimate the standard error precisely, GRF requires a large number of trees, but it will also dramatically increase the computation cost. In the Monte Carlo Simulation section, it is infeasible. Therefore, all the calculations are the same based on the 2000 tree forest. In this case, we can still observe the discernible trend from the result.

From the table below, Table of Noise Variable Test For Simple Model, we can clearly observe that at the points close to the boundary, -1 and 1, the root mean square error increases significantly with the growth of the number of noise variables. For the point at -1, the growth rate is around 1.8 from 0 noise variables to 25 noise variables. for the point at 1, it has around 1.64 growth rate. Furthermore, at the point closer to the center, -0.5(around 30.85% quantile) and 0.5(around 69.15% quantile), we can see a weaker growing trend. From 0 noise variables to 25 noise variables, the growth rate becomes around 0.4. Taking a closer look at the table, we can

Table 1: Table of Noise Variable Test For Simple Model

Criteria	Covariates	0	5	15	25
RMSE	$P_1 \& P_2 = -1$	0.474	0.928	1.21	1.33
	$P_1 \& P_2 = -0.5$	0.507	0.494	0.616	0.685
	$P_1 \& P_2 = 0$	0.458	0.344	0.259	0.238
	$P_1 \& P_2 = 0.5$	0.466	0.537	0.613	0.675
	$P_1 \& P_2 = 1$	0.504	0.891	1.204	1.331
Mean Standard Error	$P_1 \& P_2 = -1$	0.501	0.501	0.513	0.515
	$P_1 \& P_2 = -0.5$	0.475	0.476	0.479	0.496
	$P_1 \& P_2 = 0$	0.46	0.473	0.493	0.485
	$P_1 \& P_2 = 0.5$	0.469	0.48	0.489	0.495
	$P_1 \& P_2 = 1$	0.503	0.486	0.499	0.517
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0.936	0.562	0.338	0.258
	$P_1 \& P_2 = -0.5$	0.902	0.904	0.83	0.812
	$P_1 \& P_2 = 0$	0.916	0.976	0.992	0.994
	$P_1 \& P_2 = 0.5$	0.924	0.888	0.842	0.806
	$P_1 \& P_2 = 1$	0.9	0.608	0.338	0.26

[a] 0,5,15,25 are the number of noise covariates

observe that when both covariates are -0.5, it shows little fluctuation from 0 to 5 noise variables, but when it grows, it negatively affects the estimation again. At the center, coinciding with the guess before, it shows a strong decreasing trend, from 0.458 to 0.238. It is exactly opposite to the point at the boundary. In the solid over-fitting context, when the number of noise covariates increases, it alleviates the issue. Moreover, it not only reflects on the root mean square error but also on the confidence interval. In the boundary point, they have a relatively large coverage rate in the beginning, but it decreases as the increasing of number of irrelevant variables. In contrast, at the center, the coverage rate increases when there are more noise variables. Furthermore, we can found that it is not only the case for linear model, in the polynomial model, we can observe the same pattern. Please reference Table of Noise Variable Test For Polynomial Model in the appendix, we can again discern that the RMSE number increases dramatically at the boundary. At -1 covariate point, it has around 1.8 increasing rate, and at 1 covariate point it has around 1.64 increasing rate, which is almost exactly the same as the increasing rate in linear model. When we get closer to the center, at -5 and 5, the RMSE does not necessarily show a drop, but as the number of covariates increases, the growing pace slows down obviously.

Moreover, when the point is exactly at the center, the trend totally reverses. It drops significantly from 0.463 to 0.24 and these trend from boundary point and center point again reflects on the confidence interval coverage rate. There is a clear drop of coverage rate at the boundary, which implies it can barely cover the true treatment effect, but for the point at the center, it

maintains the coverage rate at around 99%.

On top of that, Table of Noise Variables Test For PMWIT in the appendix also exhibits a similar trend. At the boundary point, it shows a significant increase for RMSE, and at around 30% and 70% quantile points, the growing trend still exists but is slowing down. In the center, the number of RMSE dropped dramatically. Similarly, it reflects on the performance of the confidence interval. Therefore, if we go back to the general table, we can know that we actually average two opposite trends together. There is no fixed answer that increasing the irrelevant variables will bring up the RMSE or decrease it in this context. When the power from the boundary is larger, we will see the increasing RMSE trend, and when the power of solving the overfitting around the center is stronger we can see the decreasing trend. Only when the overfitting issue is solved, we can determine that an increase in the noise variables will negatively affect on models estimation performance.

Next, please reference the table Sample Size Experiment Result In General Case in the appendix. Now we know that the default settings of GRF, with our RD design model, will lead to the overfitting result for estimating the data from this data-generating process. However, for the sample size test, it may impact the points in a different way. First, we will still start from the general trend with 1000 fixed observations to see the average result. After that, we will again delve into different points of covariates to have a clear view.

In the simple scenario, we can easily observe that, in general, there is a decreasing trend of RMSE when the sample size increases. To be more specific, the linear model and polynomial model again almost have the same pattern and very similar values. For PMWIT, there are additional terms, X^2, $D * X^2$, and $D * X^2$, which again cause some effects on estimation. Therefore, PMWIT shows a slightly higher RMSE result. In terms of mean, standard error, again, "in average," it shows a decreasing trend. For medium case, we can first see that because linear model is lack of X^2 term, therefore it has a relatively high RMSE than the other two models. The PMWIT has two additional interaction terms, so it has a slightly higher RMSE than the polynomial model.

In the complex case, becasue PMWIT correctly specified the model, therefore its RMSE is lower than the other two models. Furthermore, again, because the interaction term and squared term are absent in the linear model, so it shows a higher RMSE than the polynomial model. Similarly, in terms of mean standard error, due to lacking interaction term in polynomial model, the X term in the model seems absorb some effects from treatment, therefore it has relatively low mean standard error which directly affects on confidence interval coverage rate being quite low. This issue happens across all three general tables.

Now, backing to the RMSE section in the general sample size experiment table, we can again see some fluctuation of RMSE when the sample size increases. In the medium case, from 2000 to 3000 sample size, the average RMSE for the linear model almost does not change. Furthermore, for the polynomial model, it shows a slight increase from 2000 to 3000 sample size increase, and for PMWIT, the pattern is the same. In the complex case, we can again see a similar issue. For the linear model, there is a small increase of RMSE from a 3000 sample size to a 4000 sample size, and this is also the pattern for the polynomial model and PMWIT. The reason can be explained as follows. As we know, increasing the sample size may not solve the over-fitting problem. However, by the same setting of GRF, increasing the sample size may create an over-fitting issue at the boundary point. In the benchmark case, there are 1000 observations, and for most of them, their covariates are located near the center. Therefore, the points at the boundary will experience less partitioning. However, when the sample size increases, the point close to the boundary may also be split further. Therefore, we can reasonably suspect that increasing the sample size expands the over-fitting problem from the center gradually towards the boundary. To further prove this point of view, I first used the same parameter in GRF to estimate the performance in different points on covariates in the simple scenario DGP.

Please first reference to Table of Sample Size Test For Linear Model in the appendix. Here, we can see that, without surprise, there is some instability and fluctuation at the center point as the increasing of sample size. But when we move outward, we can see that -0.5 and -1 also show some fluctuations. For the 95% confidence interval coverage rate, some of them also struggle at 0.9 and have fluctuation. Furthermore, this is also the case for polynomial models and polynomial models with interaction terms. Please reference Table of Sample Size Test For Polynomial Model and Table of Sample Size Test For PMWIT in the appendix. Similarly, from these two tables, we can observe that it is not only the situation for the point near the center, but it may also have a chance to happen at the point closer to the boundary. On top of that, if we check the mean standard error and confidence interval coverage rate in these two tables, we can also observe a similar result as the linear model table. However, these tables only show that the instability issue expands to the boundary, but all of the points that we test on the covariates show the fluctuation in terms of RMSE and no specific trend about confidence interval coverage rate. Therefore, there is no comparison to show how much the over-fitting problem affects the model in terms of different points and different models as the sample size increases.

To make the result clear, three additional tables have been made to use the pruned forest for estimation. In the pruned forest, a restriction to prevent unbalanced splitting has been added,

which is based on all the variables in the RD design model. Furthermore, the minimum node size requirement has been raised to 50. But because the tree is honest tree, the splitting observations and estimating observations are different, therefore, it may still have the case that there are node having the value below the requirement. From the table below, Sample Size Test For Linear Model in Pruning Forest, we can notice that pruning the tree largely boosts precision at all points in different levels of sample size. 0.2 or 0.1 is unprecedented in the previous sample size general table and sample size point testing table. To be more specific, in the RMSE section, at the point closer to the covariate boundary, -1 and 1, it shows a dramatic improvement from 1.217 to 0.232 and from 1.207 to 0.243. Also, from the confidence interval coverage rate, we can see a very clear growth up trend at the boundary point. From around 0 coverage rate to 0.728 coverage rate. Based on this information, we can expect that if the sample size increases further, the estimation at the boundary still has a higher chance of continuing to progress and allowing RMSE below 0.2.

Table 2: Sample Size Test For Linear Model in Pruning Forest

Criteria/N Size	**Covariates**	**1000**	**2000**	**3000**	**4000**
RMSE	$P_1 \& P_2 = -1$	1.217	0.675	0.401	0.232
	$P_1 \& P_2 = -0.5$	0.244	0.171	0.228	0.214
	$P_1 \& P_2 = 0$	0.233	0.243	0.21	0.18
	$P_1 \& P_2 = 0.5$	0.258	0.168	0.238	0.215
	$P_1 \& P_2 = 1$	1.207	0.691	0.397	0.243
Mean Standard Error	$P_1 \& P_2 = -1$	0.167	0.141	0.134	0.137
	$P_1 \& P_2 = -0.5$	0.168	0.174	0.209	0.207
	$P_1 \& P_2 = 0$	0.256	0.26	0.224	0.196
	$P_1 \& P_2 = 0.5$	0.166	0.171	0.207	0.2
	$P_1 \& P_2 = 1$	0.168	0.14	0.134	0.138
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0	0.002	0.236	0.728
	$P_1 \& P_2 = -0.5$	0.816	0.938	0.884	0.92
	$P_1 \& P_2 = 0$	0.956	0.928	0.936	0.956
	$P_1 \& P_2 = 0.5$	0.758	0.934	0.882	0.918
	$P_1 \& P_2 = 1$	0	0.004	0.214	0.662

Next, for the points 0.5 and -0.5, we can see that with this level of pruning, they have relatively precise performance at around 2000 sample size, which reaches RMSE 0.17 and 0.16, and it is much lower than the case without pruning. There, they have RMSE of 0.46 and 0.47, respectively. However, when the sample size increases further, the over-fitting issue reemerges due to insufficient pruning levels. It also reflects the issue on the difficulty of confidence interval progressing. For the center point, from the beginning, the level of pruning is insufficient, so the

RMSE is still above 0.2 in most of the cases. Consequently, we can say it is reasonable that the RMSE of the center point shows a fluctuation pattern as the sample size increases. Nevertheless, the RMSE value still shows an improvement in all the sample size values when compared with the previous unpruned model.

In this analysis, an important observation emerges when examining a sample size of 1000. Some point yields a relatively precise estimation that the confidence interval has above 95% coverage rate. However, with the same estimation parameters, covariates, and models, some points nearby the boundary have a coverage rate close to zero. This shows that knowing the distribution of the original data-generating process and locating the estimating point is surprisingly important for estimation. According to the point, the number of sample sizes may require a very different parameter setup for the GRF model. Additionally, we can also say that after alleviating the over-fitting problem, we can observe that as the sample size increases, RMSE will drop, and the confidence interval can reach a 95% level.

In the polynomial model, we can observe that the pattern is similar to a linear model. For the details please reference the table in appendix Table of Sample Size Test For Polynomial Model In Pruning Forest The point close to covariate boundary point, -1 and 1, they have an obvious decrease trend from around 2 to around 0.4. However, compared to the RMSE level in the pruned forest, 0.4 is still too large, which implies that the level of pruning is still too strong for the polynomial model at the boundary point when the sample size is 4000. It also reflects on the confidence interval; the coverage rate tends to zero, again, showing that the estimation is still far from the average level. Differently, in the linear model, at the -1,1 point, although the coverage rate is still low, at least it reaches around 70% in the 4000 sample case. This may imply that the linear model has a stronger over-fitting problem than the polynomial model. In the inner point around -0.5 and 0.5, again, this level of pruning may only be suitable for the sample size between 2000 and 3000. Both of them reach a relatively low RMSE at 0.14. When the sample size increases, the over-fitting problem reemerges.

Next, please reference the table Sample Size Test For PMWIT in Pruning Forest below this paragraph. From the result of the pruning polynomial model with an interaction term(PMWIT), we can discern that this level of pruning makes the result far from the regular level at the boundary point. The RMSE is around 0.6 for -1 and 1 at 4000 tree context, which is still much larger than the standard level for pruning trees 0.2. Additionally, this level of pruning appears to be appropriate for the 4000 points case at both 0.5 and -0.5 points; both of them reach the lowest point in the sample size growing up process with no fluctuation observed in between. This contrasts with the polynomial model and linear model at the same points. Therefore, we

can also conclude that PMWIT might exhibit the lowest over-fitting level in this data-generating process. Theoretically, we can say it is because PMWIT has more terms in the local regression model, and all their gradients need to be taken into account in the splitting process. Therefore, when doing the splitting, the algorithm not only needs to consider the treatment but also consider the running variables, squared terms of running variables, and other terms. Therefore, it leads to the result that PMWIT shows a relatively less over-fitting compared with other models.

Table 3: Sample Size Test For PMWIT in Pruning Forest

Criteria/N Size	**Covariates**	**1000**	**2000**	**3000**	**4000**
RMSE	$P_1 \& P_2 = -1$	1.999	1.207	0.89	0.66
	$P_1 \& P_2 = -0.5$	1.00	0.241	0.168	0.154
	$P_1 \& P_2 = 0$	0.174	0.184	0.168	0.166
	$P_1 \& P_2 = 0.5$	1.001	0.26	0.17	0.144
	$P_1 \& P_2 = 1$	2.02	1.206	0.901	0.681
Mean Standard Error	$P_1 \& P_2 = -1$	0.181	0.153	0.137	0.131
	$P_1 \& P_2 = -0.5$	0.181	0.15	0.148	0.163
	$P_1 \& P_2 = 0$	0.181	0.227	0.2	0.187
	$P_1 \& P_2 = 0.5$	0.18	0.151	0.147	0.182
	$P_1 \& P_2 = 1$	0.181	0.152	0.139	0.131
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0	0	0	0
	$P_1 \& P_2 = -0.5$	0	0.754	0.9	0.93
	$P_1 \& P_2 = 0$	0.96	0.984	0.964	0.95
	$P_1 \& P_2 = 0.5$	0	0.704	0.898	0.958
	$P_1 \& P_2 = 1$	0	0	0	0

4.2 Second stage Monte-Carlo simulation

After completing the traditional Monte-Carlo simulation, in this part, the Wasserstein Generative Adversarial Networks(WGAN) model will be applied for the subsequent stage simulation. The main limitation of traditional simulation is that the data-generating process may be tailored for some specific methods, yielding a result that seems reasonable, but in fact, the generating process cannot reflect the real-world situation. Consequently, it affects the credibility of using traditional Monte-Carlo simulation to compare the model. For the WGAN model, the real-world data will be provided as input, then the model will mimic the original data to generate estimated data, and these data will be taken as Monte-Carlo simulation data. Based on the neural networks platform, the simulation is more fair with all the models or algorithms. Here, for RD design context, I will apply the data from Ludwig and Miller (2007a). Besides, in the appendix, we can find the extension for a more generalized case(without running variable); the

data from Connors et al. (1996) would be applied. There, we can see the capability of WGAN to generate a variety of variables, including different continuous distribution data and balanced and unbalanced categorical data. Afterward, different parameters from the Generalized Random Forest will be tested. For regression discontinuity design context, it is about the HTE estimation of the Headstart program in the U.S. for the poorest counties. For the general heterogeneous effects context, the data is for the HTE of right heart catheterization. Besides Monte-Carlo simulation, one of the objectives here is also to have basic knowledge about real-world data. Therefore, when it is applied again in the application section, we can have an insightful understanding of the background. To maintain the cohesion and coherence of the thesis, the theatrical introduction of the WGAN model will be put in an appendix, please check WGAN.

Following, we will take a look at the basic information from Ludwig and Miller (2007a) Headstart data. The table is the list of variables will be used in the neural network training process.

Table 4: Auxiliary Covariates For HeadStart Program

Variable	**Description**
mort_age59_related_postHS	Mortality, Ages 5-9, Head Start related causes, 1973-1983
povrate60	County poverty rate in 1960, percent
census1960_pop	Census 1960: county population
census1960_pctsch1417	Census 1960: % attending school, age 14-17
census1960_pctsch534	Census 1960: % attending school, age 5-34
census1960_pctsch25plus	Census 1960: % high-school or more, age 25+
census1960_pop1417	Census 1960: population, age 14-17
census1960_pop534	Census 1960: population, age 5-34
census1960_pop25plus	Census 1960: population, age 25+
census1960_pcturban	Census 1960: % urban population
census1960_pctblack	Census 1960: % black population
State	State in America

Head Start is a program specifically targeted at families with children aged 0 to 5. It offers comprehensive services. From the learning perspective, they have early-age children learning and development courses. In terms of health, they prepare nutritious breakfasts and lunches for children. On the other hand, they take care of the mental health status of parents and kids to ensure a healthy upbringing. On top of that, they also support the family by covering the cost of visiting a doctor and helping the children who have special needs. Besides, the program also provides parenting support and prenatal care and hosts activities to enhance the bond between children and parents. Other than that, they provide parenting courses and facilitate meetings for parents so that they can share and learn from each other. In total, their funding amount is

about 7 billion US dollars. Our simulation and application would focus on the treatment effects of the Headstart program on health perspectives. To be more specific, the research aims to observe how the program affects the mortality rate for kids aged 5-9 per 100,000 persons in Headstart program-related causes. The running variable is the county poverty index, and the data is county-based data. The cutoff point is 59.1984. The counties above this threshold are eligible to receive support from the Headstart program. When OEO(Office of Economic Opportunity) launched this program they also found 300 assistance to the 300 poorest counties to help them for application. In the end, the counties that received the treatment effects are also 300 out of 2810 total observations. Additionally, the program was conducted in 1965, and the poverty data index was based on 1960 census data. And when 1960, ideally no one can expect that this program would start five years after. Therefore, running a variable manipulation problem, which was proposed by McCrary (2008) can be avoided. Furthermore, Cattaneo et al. (2017) proves that all the covariates are not correlated to the receiving treatment or not by plotting and using the difference in mean method. This information is important for GRF because when we apply those covariates, we do not need to worry about the case that taking those covariates in the splitting process would immediately separate control and treated observations. In this context, some leaves will include both treated and control observations with similar characteristics so that we can learn the treatment from them.

We can compare the generated and original data distribution in the histogram below. From the chart, we can observe that for continuous variables, only the poverty rate and school/university participant rate above 25 years old show a little bit of a shifted result. For other data like mortality, urban population, and black population, the simulation result is quite close to the original data. Additionally, the categorical data, such as treatment, is quite close to the original data. To take a closer look at it, the table in the appendix "Comparison of generated data and true data from WGAN" shows the mean and standard deviation comparison between true and generated data. From the table comparison table, we can observe that the county receiving the treatment has a relatively higher mortality rate in the original data. The place receiving the treatment has a lower population, lower school/university attending rate, less urban population, and more black population. All of these properties are also shown in the generated data. Also, for those which has a small mean and standard deviation, like school attending rate between ages 5 -34, the WGAN can generate them well, and for those that have a relatively large value for mean and standard deviation, for example, the total population in a different county in the US, the WGAN model can also generate them well. To have a clear view of the generated data, a total of 500,000 observations were generated, and among them, 53450 were treated data, and 446550

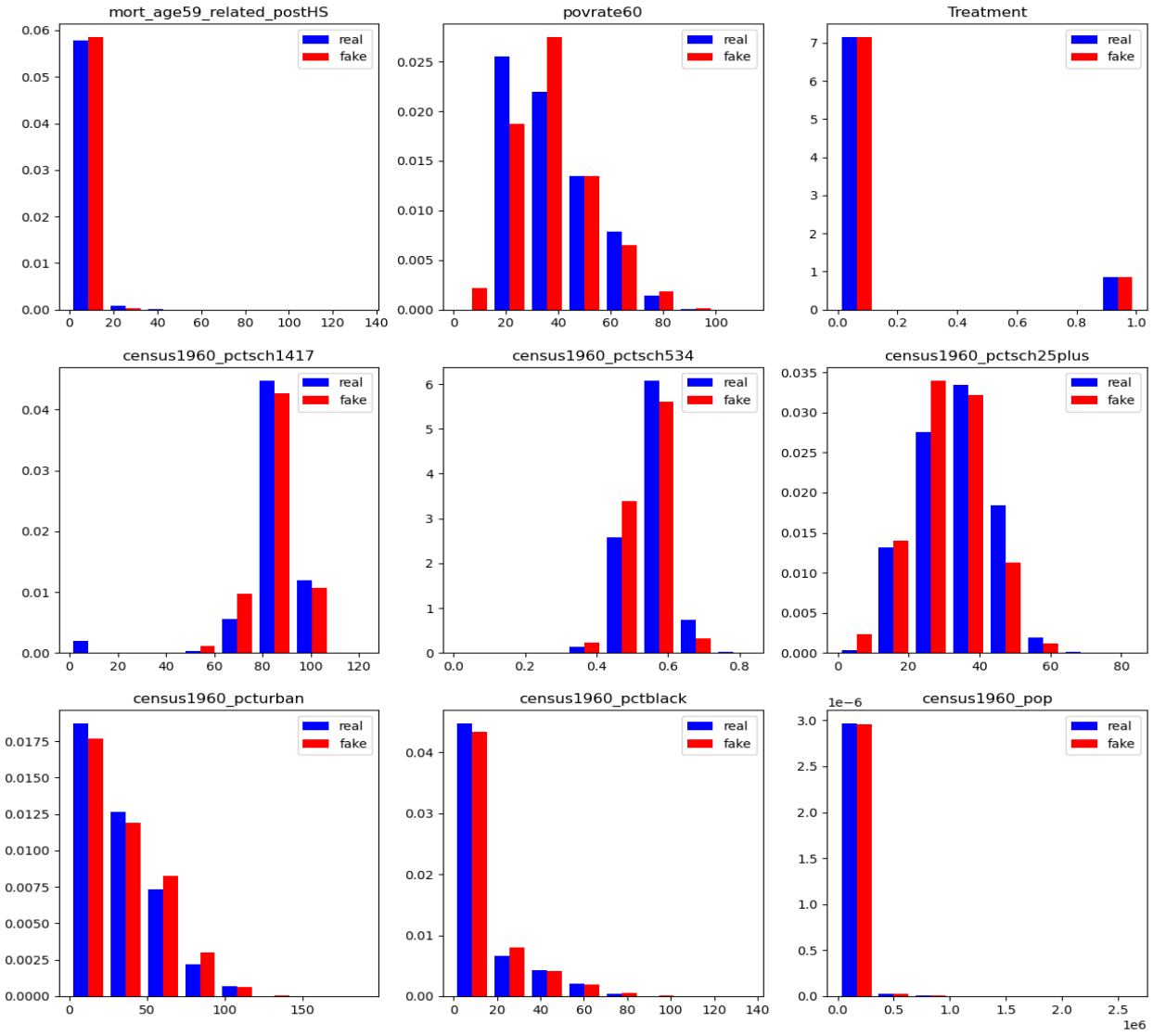

Figure 1: Continuous Variable Comparison

were controlled because the original data were unbalanced about the treated and control observations. To keep this pattern for simulating the real case, in the Monte-Carlo simulation, each time when I draw from the generated data, 100 treated data would be drawn, and 900 control data would be drawn, and two datasets would be merged together. This simulation would be iterated 500 times. Again, the linear model, polynomial model, and PMWIT will be applied in the simulation. Also, there are some ”hypothetical” observations, which are located at the 25% quantile, median, and 75% quantile of all the covariates points will be estimated. The reason for fixing the covariates is because, in this case, the simulation can coincide with our Gaussian nor-

mality result. Second, different points can have significantly different estimation results that we learn from the typical Monte-Carlo simulation; therefore, it is worth checking them separately.

Table 5: Second Stage Monte Carlo Simulation Result_Min Node Size = 5

Criteria	**Model**	**25% Quantile**	**Median**	**75% Quantile**
RMSE	Linear	1.691	1.227	1.793
	Polynomial	1.903	1.543	1.943
	PMWIT	35.75	45.26	45.38
Mean Standard Error	Linear	1.329	1.849	2.143
	Polynomial	1.648	2.289	2.556
	PMWIT	50.45	90.28	97.18
Confidence Interval Coverage	Linear	0.89	0.96	0.99
	Polynomial	0.92	0.97	0.99
	PMWIT	0.98	0.99	0.99

From the table Second Stage Monte Carlo Simulation Result_Min Node Size = 5 We can found that the estimation of PMWIT is not in the regular level; the reason behind this may be because the interaction term or squared interaction term largely affects the real treatment effect estimation(maybe absorbing the real treatment effect). Other than that, we can also observe that the result from the median is slightly better than 25% quantile and 75% quantile for both linear model and polynomial model. This may result in the quantity of data near the midpoint being relatively large and more informative. Corresponding to the real-world data background, estimating the point closer to the boundary can be a challenging task due to imbalance. Ad-

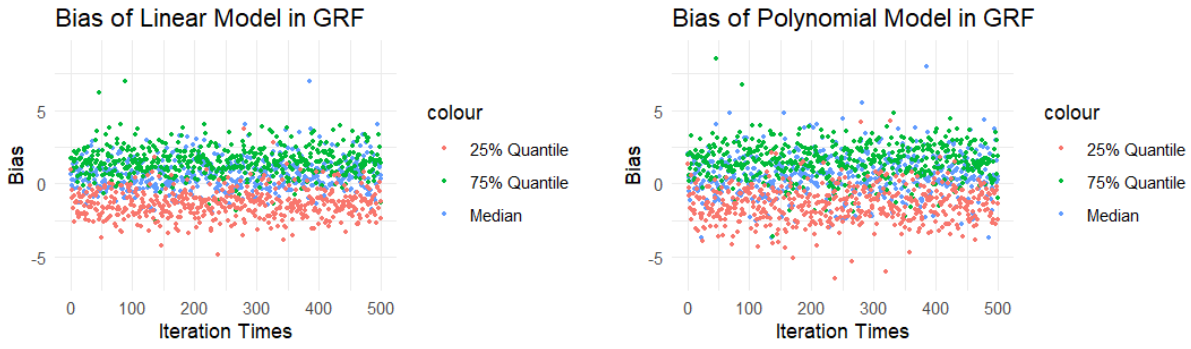

Figure 2: Estimation Bias

ditionally, when we check the bias plot of the linear model of 500 times simulation, we can

observe that the covariates point at 25% quantile usually have downward bias and covariates at 75% quantile have upward bias, and the median covariates have relatively low bias. This is also the pattern for polynomial forest. This again coincides with our expectation about the point closer to the boundary in the RDD context.

In addition to this information and analysis, there are still some other anomalies. When we check the mean standard error in the median, we can find the mean standard error is larger than the RMSE for both methods. In the 75% quantile, a similar issue happened again. Ideally, the mean squared error should be comprised by squared of bias and variance. Also, we know that the estimated standard error should be very close to the standard deviation of estimation. However, the cases I mentioned have an even larger mean standard error than RMSE. Therefore, it is reasonable to suspect that the mean standard error is overestimated. There is no problem with the RMSE value, but with this level of RMSE, it may have a lower coverage rate instead of the current confidence level in median 0.96 0.97 and in 75% quantile 0.99 and 0.99. These values may not reflect the true situation of the coverage rate. The reason behind over-estimating of mean standard error can be complicated, it can result from that misspecification of the models, because of the unbalance of data or because of over-fitting.

In the following section of the simulation, the original dataset's imbalance will be maintained because it reflects the true data conditions. Regarding model settings, two models have already been tested in the thesis and will continue to be used. Therefore, after this basic experiment about the estimation of different quantile points along the covariates on different models, the further detailed simulation will based on different settings of minimum node size and honest tree fraction. On the one hand, we test how the minimum observation settings and honest fraction settings affect our estimation; on the other hand, it may further prove our guess about the over-fitting problem in the real-world estimated dataset.

From the table, Second Stage Monte Carlo Simulation Result_Min Node Size = 15 we can observe that first, for RMSE, both local regression models show a drop at 25% quantile, median and 75% quantile point estimation.

Additionally, from mean standard error, for both 25% quantile and 75% quantile, the value is already below the RMSE. The mean standard error at the median point also closer to RMSE. Therefore, the confidence interval coverage rate may be more reliable than the tree without specific pruning. Moreover, on top of the typical simulation in the first stage, here the result again exposes the estimation issue closer to the boundary. In the context of regression discontinuity design, due to the lack of a comparison sample(they are either treated or controlled but not both), it brings difficulty in estimating the point closer to the boundary.

Table 6: Second Stage Monte Carlo Simulation Result_Min Node Size = 15

Criteria	**Model**	**25% Quantile**	**Median**	**75% Quantile**
RMSE	Linear	1.53	1.11	1.73
	Polynomial	1.767	1.301	1.844
	PMWIT	30.13	34.3	33.4
Mean Standard Error	Linear	0.984	1.398	1.359
	Polynomial	1.2	1.72	1.615
	PMWIT	36.94	64.463	58.45
Confidence Interval Coverage	Linear	0.69	0.974	0.82
	Polynomial	0.73	0.974	0.848
	PMWIT	0.964	0.988	0.992

In the next, Second Stage Monte Carlo Simulation Result_Min Node Size = 25, I further increase the minimum size requirement to 25 observations. Same as the pattern we have before, all the values in the RMSE show an obvious decrease. Also, the mean standard error has fallen to a relatively normal level. The result also coincides with our guess that it has an over-fitting problem, especially at the median. After two times of pruning, the result has a discernible improvement from RMSE, and the 95% confidence interval coverage rate is maintained at 0.96 and 0.97 with a more reasonable estimated standard error.

Table 7: Second Stage Monte Carlo Simulation Result_Min Node Size = 25

Criteria	**Model**	**25% Quantile**	**Median**	**75% Quantile**
RMSE	Linear	1.468	1.06	1.726
	Polynomial	1.69	1.203	1.796
	PMWIT	28.74	33.63	30.93
Mean Standard Error	Linear	0.872	1.166	1.17
	Polynomial	1.07	1.481	1.325
	PMWIT	31.54	54.733	46.87
Confidence Interval Coverage	Linear	0.638	0.962	0.73
	Polynomial	0.688	0.97	0.776
	PMWIT	0.96	0.988	0.98

Besides limiting the min node size of the tree, modifying the fraction of honest trees may also solve the overfitting problem. In the original simulation, the fraction of an honest tree is 0.5 for splitting and 0.5 for estimation. In the further simulation, I test 0.3 and 0.2 fractions for splitting and 0.7 and 0.8 for estimating. If there are fewer observations joining the splitting process, this implies that there is less partitioning for each tree, and more observations fall into

the leaf. Through this method, we can limit the growth of trees and, at the same time, decrease the mean standard error. From the table in the appendix Second Stage Monte Carlo Simulation Result_Honest Fraction = 0.2 and Second Stage Monte Carlo Simulation Result_Honest Fraction = 0.3 we can observe that the result exactly meet our expectation. There is a drop for both RMSE and mean standard error for both polynomial and linear models as the fraction drops. This result is similar to the pattern of the simulation of min node size. Also, the model does not have a precise estimation at the boundary point. In contrast, at the median, the confidence interval performance can be maintained at around 95%. Simultaneously, the mean standard error is lower than RMSE.

5 Application of GRF on Heterogeneous Treatment Effects Estimation

In the application part, the dataset used in the Monte Carlo simulation will continue to be use here.

5.1 Application:Headstart Program

The following are short notes about the preprocessing of the data. First, I found that many values were missing in the original dataset. Therefore, when doing the application, I replace them with the mean of the corresponding column instead of deleting them because the dataset is not large, so the rest of the data should be saved. Also, I kept the data point that has the missing value on running variables(replace them with mean), so I have six more data points than Cattaneo et al. (2017). From the simulation part, we know that the PMWIT model is not suitable for the dataset, therefore, I limiting the analysis based on linear model and polynomial model. Moreover, from the initial estimation result, I found that the confidence interval for the linear and polynomial models is much larger than the result from Cattaneo et al. (2017), and it may not be a reasonable value. Therefore, based on the previous simulation research, pruning can be an essential and important alternative for estimation results. Therefore, to have a more stable estimation, first, I have a linear pruned tree with 25 minimum node size requirement and also a polynomial pruned tree with the same requirement. For comparison, I also applied the pruning tree by adding the restriction of the unbalanced split. However, because the original dataset has 300 out of 2810 counties that received the treatment, the effects of unbalanced restriction may lead to over-pruning results. Furthermore, I also applied optimal MSE bandwidth mentioned in

Imbens and Kalyanaraman (2011) with triangular kernel also for comparison purposes. All the results can be seen in the following table

Table 8: Estimation Result Comparison

GRF Estimation			Comparison Estimation		
Model	**ATE**	**CI**	**Model**	**ATE**	**CI**
Linear_25	-1.5	[-4.79, 1.79]	Linear Flexi_h=9	-1.89	[-3.82, 0.03]
Poly_25	-1.53	[-5.79, 2.73]	Linear Flexi_h=18	-1.19	[-2.5, 0.1]
U_R_Linear	-1.41	[-3.57, 0.74]	Quartic_h=20	-2.75	[-5.5, -0.01]
U_R_Poly	-1.27	[-3.67, 1.21]	Constant R_h=3.2	-2.11	[-4.96, -0.15]
Linear_h = 6.8	-2.10	[-7.03, 2.82]	Linear R_h=6.8	-2.40	[-5.46, -0.1]
Poly_h = 6.8	-2.13	[-6.68, 2.41]	Linear R_h=9	-2.18	[-5.77, -0.35]

25 stand for the pruning level is minimum node size 25
U_R_ stand for unbalanced restriction
Flexi means flexible parametric RD methods
h is the Bandwidth the corresponding model applied
R stand for Robust bias-corrected methods
Default Linear Forest, ATE = -1.54 and CI = [-7.32,4.24]
Default Polynomial Forest, ATE = -1.86 and CI = [-10.45,6.72]

From the estimation in Ludwig and Miller (2007a) instead of using the bandwidth selection methods, they specified an ad-hoc bandwidth for estimation, like in the comparison section h=9 , h= 18 and h = 20 in the table. To replicate this result, from the paper, Cattaneo et al. (2017) calculate the difference between running variables and cutoff point. Based on the difference and ad-hoc bandwidth, they build up a regression model for estimation. Besides the ad-hoc bandwidth specifying method, they also applied non-parametric methods. First, apply optimal MSE methods for bandwidth selection. Second, in the local regression model, they applied kernel to weight the different distance between running variables and the cut-off point. In the triangular kernel, the point closer to the cut-off point will obtain a higher weight. Based on this weighting result, Cattaneo et al. (2017) build up a typical non-parametric regression discontinuity model. In the table, Constant R, Linear R are from this method.

Comparing the results, we can notice that the estimation of the average treatment effects from the generalized random forest is lower than that of the typical non-parametric RDD method. Also, we can see that for the polynomial and linear models, the confidence interval is slightly larger than all the results from Cattaneo et al. (2017). The reason behind this can be that the entire dataset is applied in the generalized random forest instead of only using the observations inside the bandwidth.

Therefore, the outliers or the point closer to the boundary may make the estimation result more diverse. Also, the goal of the splitting criteria is to make the treatment effects heteroge-

neous. Therefore, it is reasonable that it has a larger standard error estimation, which leads to a larger confidence interval. To understand it from the perspectives of the dataset, it can be observed that in the counties with low school attendance and suffering from poverty if they receive the treatment, the treatment effects for them can be enormous. However, because of the high poverty index, they may be excluded from the consideration of the typical regression discontinuity model. In contrast, for those places that have low poverty index and high school attendance rate, the treatment effects for them will be relatively low. Those are also not able to be included in traditional RD design models. Moreover, in the original dataset, as we mentioned, only 300 out of 2810 counties receive the treatment effects. So among these 2510 control observations, many of them may have relatively low treatment effects from the generalized random forest estimation, these properties can also be seen from the histogram we plot in the second stage Monte-Carlo simulation. Consequently, when we take the average of heterogeneous treatment effects, the result may be lower than the typical regression discontinuity design.

Additionally, by limiting the minimum observations in the node, we can see a reasonable improvement in performance. Before pruning, the result of the confidence interval was [-7.32,4.24] and [-10.45,6.72], respectively, for linear trees and polynomial trees. But after pruning, the confidence interval shrinks to [-5.79, 2.73] and [-4.79, 1.79], respectively; on top of that, the average treatment effects do not exhibit a considerable change. This situations also emerge on the unbalance restriction linear forest, unbalance restriction polynomial forest, bandwidth linear forest and bandwidth polynomial forest.That the reason that the pruned forest has a small confidence interval is because there are more data points in the leaf, which lowers the difference between the trees in the forest. But for the bandwidth tree, it just simply consider less observations. Therefore, in terms of heterogeneity, bandwidth forests are likely to be more heterogeneous. This is also evidence in the table that bandwidth linear forest and bandwidth polynomial forest have larger confidence intervals than other pruning forest. For the later heterogeneous discussion, we can see the situation from the graph.

Next, the graph of 25 minimum node size pruned linear forest and 25 minimum node size pruned polynomial forest has been plotted. Please reference the appendix Linear_25 Forest and Poly_25 Forest Because the size of the structure of the tree, they are put in appendix. Also, it may be the case that it is still not clear enough in the appendix. Please check the footnote, from the link, the .svg graph will be shown, and it can be zoom in. All the tree structure from different models are put there.

When comparing the outcome of 25 min node size pruned tree and the unbalanced restriction pruned tree, we can notice that the 25 min node size pruned tree is more complex for both the

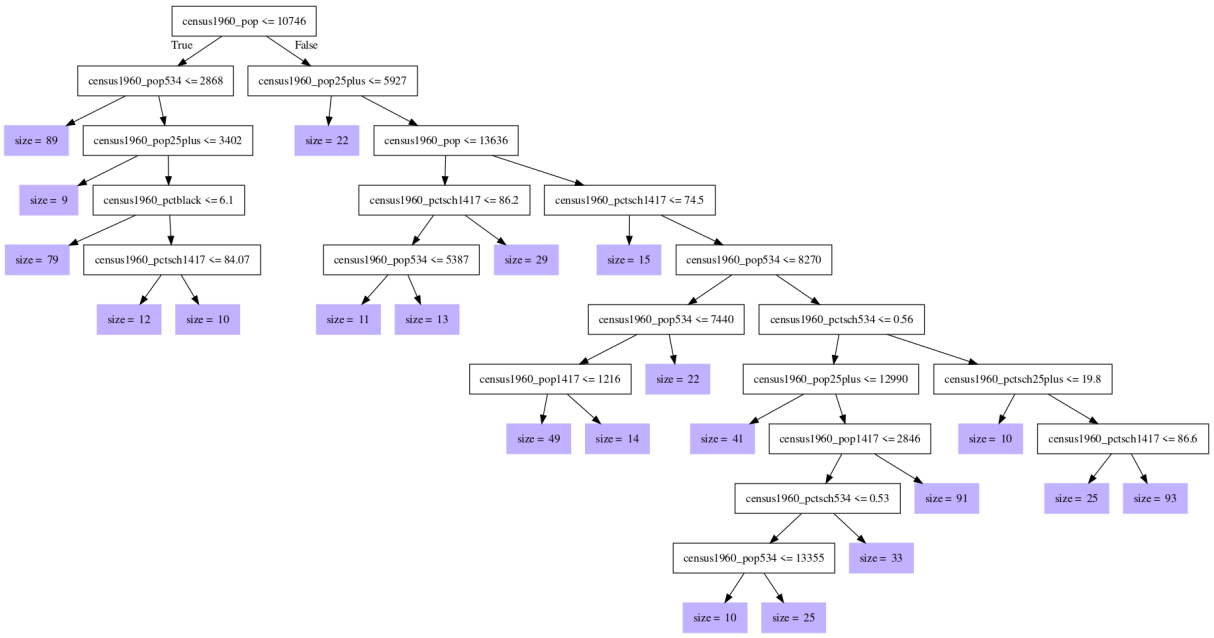

Figure 3: One of tree from unbalance restriction linear forest

linear model and the polynomial model than the unbalanced restriction pruned tree. The reason can again be because the original dataset is already unbalanced. In the unbalanced context, it is more challenging to prevent the unbalanced treatment effects term, running variable term, and square of running variable term be allocated unbalanced. Therefore, from the pruned tree structure, we can observe that many of the branches stop in the relatively early phase. There are even some leaf end up at around 90 observations. Consequently, this may leads to a less heterogeneous result, because the affects may be ”average away”. Besides, the unbalanced restriction polynomial tree also shows a similar form, and it can also be found in the appendix titled UR$_p$*oly*.

Because the size of the pruned tree is too large, please either check in the appendix or in the footnote link[1] Next is the tree be constructed only by the dataset inside the Imbens-Kalyanaraman MSE-optimal bandwidth, other than that, there is no specified pruning method be adopted for these two bandwidth forest.

The graph below exhibit the structure of linear forest with optimal MSE bandwidth. We can observe that the level of complexity of the tree is similar to the unbalanced restriction tree. However, differently, it can keeps the obvious heterogeneous effects in the forest, by without doing further specific pruning.

In the next paragraph, to observe the heterogeneous effects in details, the distribution of heterogeneous effects will be plotted.

Firstly, from the histogram,Heterogeneous Treatment comparison 1 the linear forest with

[1]Clear version:GRF Tree Structure

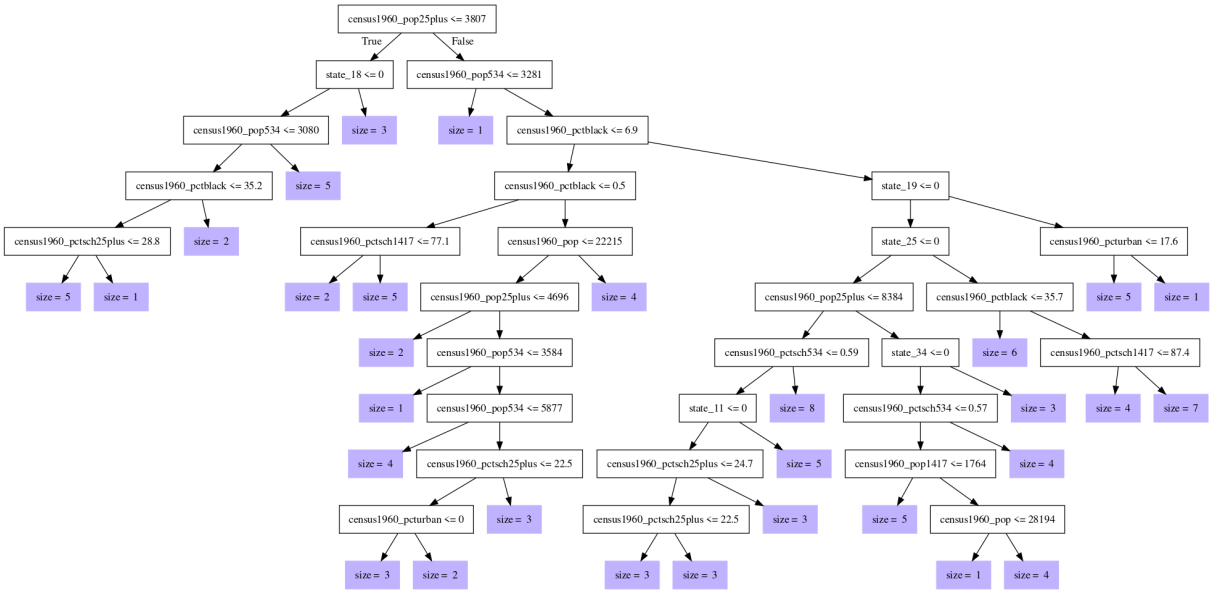

Figure 4: One of the tree from bandwidth linear forest

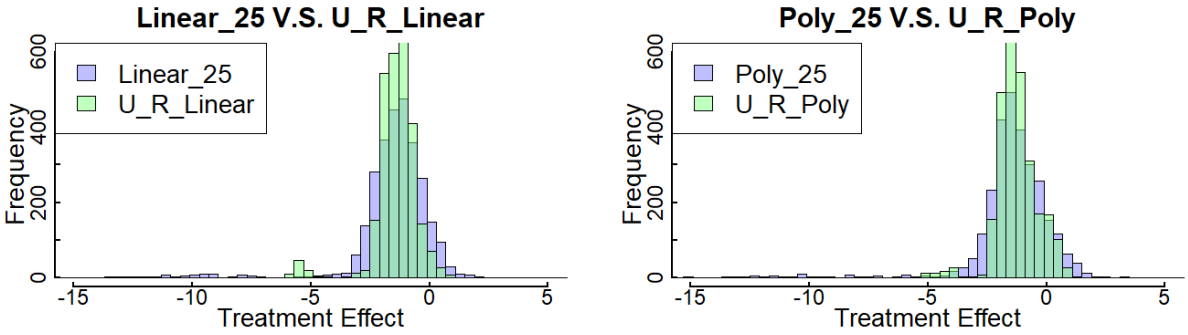

Figure 5: Heterogeneous Treatment comparison 1

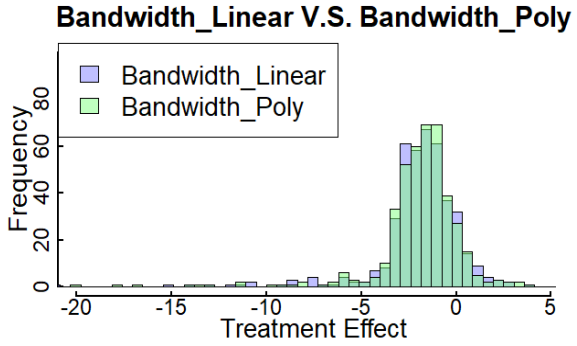

Figure 6: Heterogeneous Treatment comparison 2

25 min node size is more heterogeneous than the imbalance restriction forest, and the majority of the treatment effects are located between -5 and 0. Other than that, we can see a long tail to the left that can almost reach 15, showing the various different levels of treatment effect, particularly from the 25 min node size linear model. From the graph on the right-hand side, we can observe that the polynomial forest with 25 min node size is, again,more heterogeneous than the forest with unbalanced restriction, which coincides with the table that the forest with node size restriction has a slightly wider confidence interval. Also, from the plot of the tree, we observed that an unbalanced restriction forest accumulated more observations in the leaf, which also affects the distribution, making it located more central. Next, we go to the graphs illustrating the distribution of bandwidth forest. Noting that because it only takes into account the points inside the bandwidth, the plot can hardly be plotted together with the other two genres of the forest. Therefore, they are plotted separately to show the details. From the graphs, Heterogeneous Treatment comparison 2, we can see it has a relatively wider distribution; from the left, it can reach around 20, and from the right, it can reach almost 5. In terms of number, the smallest number bandwidth forest can reach is around -26; for the min node size restriction forest, it can reach around -18, and for the unbalanced restriction forest, the smallest treatment effect is just below -5. These numbers are the mortality rate per 100,000 children caused by the Headstart program specified.

Additionally, we can see that there are proportionally fewer individuals located close to zero treatment effects compared with node size restriction forest and unbalanced restriction forest plots in terms of distribution shape. This can coincide with our guess that it may be because the majority of observations eliminated from the bandwidth are those treatment effects close to zero. In the dataset, only 300 observations receive the treatment affects. Some of the counties that have very high school attendance rates, large populations, and are more urbanized may be eliminated from the model. Consequently, we can observe from the bandwidth forest graph that the amount of observations located around -2.5 is not very different from those located around -1.

To conduct a comprehensive study of heterogeneous effects on generalized random forest. An extension of the application based on the context without running variables would be done and put in the appendix for the cohesion of the main text. There is crucial idea will be introduced there: Rank weighted average treatment(RATE) effects, proposed from the paper Yadlowsky et al. (2021). In the entire thesis, all the estimations are heterogeneous based. However, we still do not have a method to quantify our target.In 2021, Yadlowsky et al. (2021) present the method RATE, which can not only quantify the level of heterogeneity but also estimate the

significant level of heterogeneity. To comprehensive the research of GRF on regression discontinuity design, it is important to cover this method. However, because generalized random forest is still a relatively new method, the R-package grf do not support multivariates local regression models for RATE estimation. Therefore, in the following application, the equation1 model will be applied so that we can observe the RATE result.

6 Conclusion

The goal of the thesis is aiming at theoretically and applicably build up the regression discontinuity models on generalized random forest framework for heterogeneous treatment effects estimation. The contribution of the thesis can be divided into two sections: theory and application. Furthermore, to bridge the gap in between, first and second stage of Monte Carlo simulation has been applied. To be more specific, in the first and second chapters, comprehensive and fundamental assumptions for building the regression discontinuity model on the generalized random forest have been established. These are mainly based on the research in Athey et al. (2019), and Wager and Athey (2018), Cattaneo et al. (2017), Imbens and Lemieux (2008), but is tailored for our special research case. The critical assumption underlying our regression discontinuity design case is that we need to guarantee the potential outcome will not exhibit any jump by changing X when conditionally on any covariate P. This ensures that the effect being estimated is solely due to the only treatment at the cutoff point. The main difference between this assumption and the typical regression discontinuity assumption is this assumption needs to be held in the entire dataset instead of just around the cut-off point so that GRF can efficiently apply every data point to capture the pattern of heterogeneous treatment effects. Moreover, the relationship between covariates and running variables is also being regulated to ensure that in the GRF estimating process, the splitting process will not create an extremely unbalanced partition. Additionally, in terms of the assumption for GRF, to apply the core result from Wager and Athey (2018), the Lipschitz continuous assumption in the expectation of our "loss function" ψ conditionally on covariates p has to hold. Furthermore, since our interest variable in the RD model is differentiable, there are several original assumptions from Athey et al. (2019) that are no longer unnecessary. Given these adjustments and the characteristics of RD design models, some technical proofs have been simplified. As a result, from these simplified proofs, we have derived the Gaussian normality property for our regression discontinuity models on the GRF framework.

The asymptotic theorem is typically based on infinite sample size. To fill the gap between

theory and application, Monte-Carlo simulation was utilized in the context of finite sample sizes. This allows us to examine the properties and potential issues associated with certain situations. From the simulation, we can see a clear trend that without the over-fitting issue, as the sample size increases, our estimated parameters are closer to the true parameter. Also, as the noise from the data-generating process increases, it will impact the precision of the estimation. As the number of noise variables increases, without over-fitting, the root mean squared error is also brought up. All of these results can coincide with the theory. However, when there is an overfitting problem, these trends are not necessarily the case. We learn that since the covariates both follow the normal distribution without particular pruning, there is a strong over-fitting problem at the center point, and it diminishes as we move from the center toward the boundaries. Consequently, as the number of noise covariates increases, there is a noticeable decreasing trend in RMSE at the center with a significant improvement in confidence interval coverage. In contrast, due to fewer individuals located at the boundary, an increase in the number of noise covariates significantly raise the RMSE at the boundary points, this situation also reflects on the huge decline of confidence interval coverage rate.

In the sample size experiment, again, we face an overfitting problem, but in the sample size simulation context, it not only occurs at the center but also extends to the boundary as the sample size increases. To prove our guess, a detailed test along different points on covariates is conducted. From the test result, we can see that as the sample size increase, most of the points are immerse in the over-fitting issue, even the boundary points shows a struggling in terms of progression on decrease RMSE and increase confidence interval coverage rate. To address this issue even more clearly, the same data-generating process is applied, but the models are significantly being pruned. From the result of the pruning forest, we clearly observe that at the boundary point, as the sample size increases, the RMSE decreases significantly, and the confidence interval coverage rate also increases. For points between the boundary and the center, as the sample size increase, at some point the unprecedented low RMSE appeared. However, as the sample size keeps increasing, the over-fitting problem re-emerges. For the point at the center, with this level of pruning, it still has the overfitting problem from the beginning stage. From these experiments, we can clearly observe that not only which covariates being selected for splitting is important, but which point we want to estimate is also surprisingly crucial. Utilizing the same covariates but estimating different points for research interest requires distinct parameter setups. Even with the identical level of pruning applied to same covariate, the confidence interval coverage may vary obviously. At some points, it could approach nearly zero, while at others, it may be maintained at 0.95 level. Besides,the sample size can be an important

feature that we need to take into account; the increasing sample size may require a different level of pruning method to maximize the benefit brought by the sample size. Furthermore, when comparing the level of over-fitting issues on different models, we can also find that the polynomial model with interaction terms is less over-fitting. It is because PMWIT needs to consider more independent variables than other models in the partitioning phase. In conclusion, achieving precise and consistent estimations requires careful consideration of several factors, including the distribution of covariates, the location of estimation points, the sample size, and the model settings. Based on all of these issues, we can decide which level of pruning forest to apply and to prevent the result from being misleading. In the second stage, we are one step closer to the application. The real-world data from the Headstart program is applied for data generation. Based on the parameters batch size = 128, the architecture 128-128-128, with max epochs 5000 the covariates data was generated. Based on the generated covariates data, the WGAN further generates the outcome variables and the counterfactual outcome variables. Next, we evaluate the result from different perspectives. From the comparison histogram and table, we learn that in terms of distribution, mean, and standard deviation, the generated data is very close to the original data. Based on these generated data, the second stage Monte Carlo simulation is conducted. First, according to the result of the first stage of simulation, we know that it is essential to estimate the result based on specific points; therefore, in the second stage, we apply 25% quantile, median, and 75% quantile for simulation—also, three models, linear model, polynomial model and PMWIT. The results show that the linear model and polynomial model are more precise and consistent than the PMWIT for the Headstart dataset. However, the over-estimates standard error for some models at different points shows a potential problem of the estimation. The confidence interval coverage rate may not reflect the true performance of the estimation. As a result, I did another four simulations according to different min node sizes and honest fractions. Consequently, after different levels of pruning, four of them all lead to an improvement of RMSE. Also, the mean standard error are decreased to the normal level so that the confidence interval shows a more reliable result. Moreover, the result again exposes the potential problem of GRF estimation on boundary points in the RDD context.

In the application section, the Headstart program was applied again. There are 6 different RD design models on GRF framework has been applied for estimation, including 25 min node size linear forest, 25 min node size polynomial forest, unbalanced restriction linear forest, unbalanced restriction polynomial forest and optimal MSE bandwidth linear forest and optimal MSE bandwidth polynomial forest. 25 min node size forest and unbalance restriction forest in fact are different level of pruning. Because the dataset has a quite unbalance treatment and

control group, so it largely restrict the tree construction for unbalance restriction forest. Besides, for the optimal MSE bandwidth forest, because of the lower number of observations, the structure is close to the unbalanced restriction forest. These results are compared with the result from Cattaneo et al. (2017) with flexible parametric regression discontinuity methods and robust bias-corrected methods. In conclusion, we can observe that the result from GRF with RD models has a wider confidence interval and lower average treatment effects in this dataset. Among them, optimal MSE bandwidth forest shows the largest confidence interval, after that is 25 min node size forest, the last is the unbalance restriction forest. In the same order, optimal MSE bandwidth forest detects the highest average treatment effects; the value is close to the result from robust bias-correlated methods in Cattaneo et al. (2017), after that is 25 min node size forest, and then is unbalance restriction tree. When the heterogeneous effects are plotted, we can also observe that there are a significant amount of individuals who should have relatively low treatment effects excluded from the optimal MSE bandwidth forest. From the table in the second stage of the Monte Carlo simulation, we notice that the majority of individuals are not receiving the treatment effects. Therefore , when estimating the result, these groups of people lower the treatment effects for 25 node-size forests and unbalanced restriction forests. Consequently, on average, these two forests turn out to have lower average treatment effects than the optimal MSE bandwidth forests. From these perspectives, if researchers are interested in the point inside that bandwidth. The bandwidth forest can be a feasible option for estimation besides pruned forest. On one hand it filter the less informative data point, on the other hand it still maintain the heterogeneity of estimation. However, research on bandwidth generalized random forests is still relatively limited nowadays. This may result from the complexity and breadth of bandwidth selection topics. Also, because GRF is still a relatively new method, its assumptions may not be comprehensive enough to cover all the related topics. In the future, there is a clear need to construct rigorous assumptions for bandwidth GRF, so people can based on that to derive an unbiased, consistent bandwidth GRF estimator. Additionally, quantifying the effects of heterogeneous treatment is another important perspective for GRF. Based on the research McConnell and Lindner (2019), we know that GRF may not be the most precise method in terms of average treatment effects estimation. The advantage of GRF is based on its capability to capture the HTE. However, besides learning it from the graph, we need to quantify how much HTE we capture and whether there is a significant difference between leaf and leaf. A rigorous, feasible method needs to be developed and combined with GRF in a regression discontinuity context. Therefore, people can have common, standard criteria for evaluation of the result in terms of HTE estimation.

7 Bibliography

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., Imbens, G. W., Metzger, J., and Munro, E. (2024). Using Wasserstein Generative Adversarial Networks for the design of Monte Carlo simulations. *Journal of econometrics*, 240(2):105076.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of statistics*, 47(2).

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2017). Comparing Inference Approaches for RD Designs: A reexamination of the effect of head start on child mortality. *Journal of policy analysis and management*, 36(3):643–681.

Connors, Alfred F., J., Speroff, T., Dawson, N. V., Thomas, C., Harrell, Frank E., J., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, William J., J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. A. (1996). The Effectiveness of Right Heart Catheterization in the Initial Care of Critically III Patients. *JAMA*, 276(11):889–897.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5769–5779, Red Hook, NY, USA. Curran Associates Inc.

Hirano, K. and Imbens, G. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4):259–278. Funding Information: We are grateful to the SUPPORT study for making their data available, to Enrico Moretti for help working with the

data, and to Donald Rubin, Rajeev Dehejia, and three anonymous referees for comments. Sejin Min, Marcos Rangel and Yue Xu provided excellent research assistance. Financial support for this research was generously provided through NSF grants SES-9985257 (Hirano) and SBR-9818644 and SES-0136789 (Imbens).

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *Annals of mathematical statistics*, 19(3):293–325.

Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2002). Bagging survival trees. *Statistics in Medicine*, 23.

Imbens, G. and Kalyanaraman, K. (2011). 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. The regression discontinuity design: Theory and applications.

Ludwig, J. and Miller, D. L. (2007a). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. *The Quarterly journal of economics*, 122(1):159–208.

Ludwig, J. and Miller, D. L. (2007b). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design*. *The Quarterly Journal of Economics*, 122(1):159–208.

McConnell, K. J. and Lindner, S. (2019). Estimating treatment effects with machine learning. *Health services research*, 54(6):1273–1282.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714. The regression discontinuity design: Theory and applications.

MeinshausenNicolai (2006). Quantile Regression Forests. *Journal of Machine Learning Research*.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(5):141–158.

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.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects.

8 Appendix

8.1 Technical Proof For The Main Result

8.1.1 Proof of Result 1

First, we assume the following based on the convergence result of ψ function for the true parameter and estimated parameter

$$\left\| \sum_{i=1}^{n} \alpha_i(p) \psi_{\beta_1(p),\nu(p)} \right\|_2 \text{ and } \left\| \sum_{i=1}^{n} \alpha_i(p) \psi_{\hat{\beta}_1(p),\hat{\nu}(p)} \right\|_2 < \varepsilon_n$$

Second, define a series of η like the following.

$$\text{a. } \lim_{n\to\infty} \eta_n = 0 \qquad \text{b. } \eta_n > \frac{4\varepsilon_n}{\sigma^2} \forall n \qquad \text{c. } \frac{\left\| \sum_{i=1}^{n} \alpha_i(p) \psi_{\hat{\beta}_1(p),\hat{\nu}(p)} \right\|_2}{\eta_n} \to_p 0$$

Based on assumption 7, we can have the following derivation.

$$F(\beta_1,\nu) \geq F(\hat{\beta}_1,\hat{\nu}) + \nabla F(\hat{\beta}_1,\hat{\nu}) \cdot \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \nu - \hat{\nu} \end{pmatrix} + \frac{\sigma^2}{2} \left\| \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \nu - \hat{\nu} \end{pmatrix} \right\|_2^2$$

$$= F(\beta_1,\nu) \geq F(\hat{\beta}_1,\hat{\nu}) + \sum_{i=1}^{n} \alpha_i(p) \psi_{\hat{\beta}_1(p),\hat{\nu}(p)} \cdot \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \nu - \hat{\nu} \end{pmatrix} + \frac{\sigma^2}{2} \left\| \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \nu - \hat{\nu} \end{pmatrix} \right\|_2^2$$

Based on the definition of η_n, we obtain the following result

$$F(\beta_1,\nu) - F(\hat{\beta}_1,\hat{\nu}) \geq \frac{\sigma^2}{2} \left\| \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \nu - \hat{\nu} \end{pmatrix} \right\|_2^2$$

Now assume for contradiction,

$$\left\| \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \nu - \hat{\nu} \end{pmatrix} \right\|_2 = \eta_n$$

Then,

$$F(\beta_1,\nu) - F(\hat{\beta}_1,\hat{\nu}) \geq \frac{\sigma^2 \eta_n^2}{2}$$

By the gradient theorem, we obtain the following result

$$\eta_n \left\| \sum_{i=1}^{n} \alpha_i(p) \psi_{\beta_1(p),\nu(p)} \right\|_2 \geq \frac{\sigma^2 \eta_n^2}{2} \implies \left\| \sum_{i=1}^{n} \alpha_i(p) \psi_{\beta_1(p),\nu(p)} \right\|_2 \geq \frac{\sigma^2 \eta_n}{2}$$

However, this result contradicts to the definition for η_n that $\frac{\eta_n \sigma^2}{4} > \varepsilon_n$ Therefore,

$$\left\| \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \nu - \hat{\nu} \end{pmatrix} \right\|_2 < \eta_n$$

8.1.2 Proof of Result 2

First, we know that $\hat{\beta}_1$ is getting closer to β_1 when the n is getting larger from the previous result, so we can have the following assumption

$$\exists \varepsilon_n \to 0 \; when \; n \to \infty \; and \; \left\| \begin{pmatrix} \hat{\beta}_1(p) - \beta_1(p) \\ \hat{\nu}(p) - \nu(p) \end{pmatrix} \right\|_2 = O_P(\varepsilon_n)$$

Next, based on the Taylor expansion we can write the following equation.

$$\sum_{i=1}^{n} \alpha_i(p) \psi_{\hat{\beta}_1, \hat{\nu}}(Y_i) - \sum_{i=1}^{n} \alpha_i(p) \psi_{\beta_1, \nu}(Y_i)$$

$$= \left(\sum_{i=1}^{n} \alpha_i(p) \nabla \psi_{\beta_1(p), \nu(p)}(P_i) \right) \begin{pmatrix} \hat{\beta}_1(p) - \beta_1(p) \\ \hat{\nu}(p) - \nu(p) \end{pmatrix} + H$$

We can use the equation(10) again from the result of Wagner and Athey (2018)Wager and Athey (2018) because of Lipschitz continuous and the weight α.

$$\left\| \sum_{i=1}^{n} \alpha_i(p) \nabla \psi_{\beta_1(p), \nu(p)}(P_i) - V(p) \right\|_F = O_P \left(s^{-\frac{\pi}{2} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}} \right)$$

From here we know that

$$\left(\sum_{i=1}^{n} \alpha_i(p) \nabla \psi_{\beta_1(p), \nu(p)}(P_i) \right) \begin{pmatrix} \hat{\beta}_1(p) - \beta_1(p) \\ \hat{\nu}(p) - \nu(p) \end{pmatrix} = O_P \left(s^{-\frac{\pi}{2} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}} \varepsilon_n \right)$$

According to assumption 6, existence of solution, and assume C $\leq \frac{1}{\varepsilon^{\frac{2}{3}}}$, we can have the following result.

$$\sum_{i=1}^{n} \alpha_i(p) \psi_{\hat{\beta}_1, \hat{\nu}}(Y_i) \leq C \max_{1 \leq i \leq n} \{\alpha_i\} \leq C \frac{s}{n} \leq \frac{s}{n\varepsilon^{\frac{2}{3}}}$$

From here we can say that

$$\sum_{i=1}^{n} \alpha_i(p) \psi_{\hat{\beta}_1, \hat{\nu}}(Y_i) - \left(\sum_{i=1}^{n} \alpha_i(p) \nabla \psi_{\beta_1(p), \nu(p)}(P_i) \right) \begin{pmatrix} \hat{\beta}_1(p) - \beta_1(p) \\ \hat{\nu}(p) - \nu(p) \end{pmatrix} = O_P \left(s^{-\frac{\pi}{2} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}} \varepsilon_n, \frac{s}{n\varepsilon^{\frac{2}{3}}} \right)$$

Therefore,

$$\left\| \begin{pmatrix} \hat{\beta}_1(p) - \beta_1(p) \\ \hat{\nu}(p) - \nu(p) \end{pmatrix} - V^{-1}(p) \sum_{i=1}^{n} \alpha_i(p)\gamma^*(p) \right\|_2 = O_P \left(s^{-\frac{\pi}{2}\frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}} \varepsilon_n, \frac{s}{n\varepsilon^{\frac{2}{3}}}, \varepsilon^2 \right)$$

Based on the equation (10) , we can say that $(\hat{\beta}_1, \hat{\nu})$ has to be consistent with $\sqrt{\frac{s}{n}}$. Therefore, now ε is equal $\sqrt{\frac{s}{n}}$, we obtain the result, and we plug it into the original form.

$$\sqrt{\frac{n}{s}} \left(\tilde{\beta_1}^*(p) - \hat{\beta}_1(p) \right) = O_P \left(\max \left\{ s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log(\omega^{-1})}}, \left(\frac{s}{n}\right)^{\frac{1}{6}} \right\} \right)$$

8.2 Table

Table 9: Noise Experiment Result In The General Case

Criteria	RDD Model	Noise	Simple	Medium	Complex
RMSE	**Linear Model**	1	0.56	1.64	2.333
		1.5	0.757	1.721	2.335
		2	0.97	1.872	2.409
		2.5	1.206	1.98	2.526
	Polynomial Model	1	0.563	0.863	1.168
		1.5	0.759	1.061	1.284
		2	0.974	1.268	1.387
		2.5	1.209	1.445	1.536
	PMWIT	1	0.721	1.103	1.123
		1.5	1.007	1.301	1.312
		2	1.303	1.524	1.527
		2.5	1.605	1.784	1.789
Mean Standard Error	**Linear Model**	1	0.518	1.347	1.857
		1.5	0.729	1.442	1.924
		2	0.947	1.572	2.007
		2.5	1.171	1.709	2.13
	Polynomial Model	1	0.512	0.599	0.621
		1.5	0.721	0.795	0.817
		2	0.937	1.016	1.021
		2.5	1.159	1.214	1.239
	PMWIT	1	0.642	0.745	0.752
		1.5	0.924	1.005	1.015
		2	1.208	1.281	1.282
		2.5	1.50	1.572	1.564
95% Confidence Interval Coverage	**Linear Model**	1	0.903	0.871	0.861
		1.5	0.91	0.881	0.869
		2	0.911	0.879	0.870
		2.5	0.911	0.885	0.871
	Polynomial Model	1	0.9	0.883	0.753
		1.5	0.906	0.896	0.816
		2	0.906	0.902	0.846
		2.5	0.907	0.907	0.862
	PMWIT	1	0.885	0.868	0.866
		1.5	0.891	0.876	0.876
		2	0.893	0.881	0.881
		2.5	0.895	0.886	0.882

Table 10: Noise Variables Experiment Result In The General Case

Criteria	RDD Model	NV	Simple	Medium	Complex
RMSE	**Linear Model**	0	0.549	1.642	2.339
		5	0.667	1.754	2.185
		15	0.804	1.743	2.304
		25	0.881	1.856	2.424
	Polynomial Model	0	0.549	0.86	1.175
		5	0.663	1.244	1.45
		15	0.802	1.221	1.557
		25	0.878	1.3	1.662
	PMWIT	0	0.713	1.029	1.098
		5	0.839	1.566	1.579
		15	1.089	1.77	1.803
		25	1.179	1.77	1.926
Mean Standard Error	**Linear Model**	0	0.514	1.342	1.862
		5	0.468	1.23	1.742
		15	0.479	1.21	1.752
		25	0.478	1.204	1.745
	Polynomial Model	0	0.508	0.585	0.621
		5	0.467	0.585	0.621
		15	0.48	0.596	0.661
		25	0.479	0.595	0.679
	PMWIT	0	0.64	0.734	0.751
		5	0.586	0.785	0.796
		15	0.6	0.816	0.848
		25	0.604	0.834	0.868
95% Confidence Interval Coverage	**Linear Model**	0	0.907	0.872	0.859
		5	0.824	0.871	0.893
		15	0.755	0.848	0.873
		25	0.717	0.819	0.864
	Polynomial Model	0	0.903	0.881	0.737
		5	0.825	0.783	0.773
		15	0.755	0.708	0.747
		25	0.718	0.655	0.73
	PMWIT	0	0.89	0.87	0.87
		5	0.806	0.732	0.748
		15	0.689	0.668	0.655
		25	0.652	0.641	0.646

Table 11: Noise Variables Test For Polynomial Model

Criteria	**Covariates**	**0**	**5**	**15**	**25**
RMSE	$P_1 \& P_2 = -1$	0.468	0.936	1.206	1.329
	$P_1 \& P_2 = -0.5$	0.501	0.49	0.614	0.68
	$P_1 \& P_2 = 0$	0.463	0.345	0.255	0.24
	$P_1 \& P_2 = 0.5$	0.459	0.535	0.61	0.664
	$P_1 \& P_2 = 1$	0.504	0.894	1.204	1.334
Mean Standard Error	$P_1 \& P_2 = -1$	0.5	0.514	0.509	0.513
	$P_1 \& P_2 = -0.5$	0.47	0.499	0.493	0.501
	$P_1 \& P_2 = 0$	0.447	0.479	0.489	0.491
	$P_1 \& P_2 = 0.5$	0.458	0.477	0.475	0.494
	$P_1 \& P_2 = 1$	0.497	0.512	0.511	0.519
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0.942	0.59	0.342	0.232
	$P_1 \& P_2 = -0.5$	0.912	0.914	0.84	0.81
	$P_1 \& P_2 = 0$	0.896	0.974	0.998	0.996
	$P_1 \& P_2 = 0.5$	0.908	0.86	0.818	0.802
	$P_1 \& P_2 = 1$	0.924	0.642	0.35	0.246

[a] 0,5,15,25 are the number of noise variables

Table 12: Table of Noise Variables Test For PMWIT

Criteria	**Covariates**	**0**	**5**	**15**	**25**
RMSE	$P_1 \& P_2 = -1$	0.63	1.514	1.814	1.877
	$P_1 \& P_2 = -0.5$	0.628	0.844	0.945	1.011
	$P_1 \& P_2 = 0$	0.596	0.496	0.325	0.316
	$P_1 \& P_2 = 0.5$	0.613	0.881	0.974	0.982
	$P_1 \& P_2 = 1$	0.67	1.399	1.765	1.847
Mean Standard Error	$P_1 \& P_2 = -1$	0.5	0.514	0.509	0.513
	$P_1 \& P_2 = -0.5$	0.597	0.675	0.67	0.677
	$P_1 \& P_2 = 0$	0.58	0.666	0.675	0.68
	$P_1 \& P_2 = 0.5$	0.588	0.656	0.662	0.670
	$P_1 \& P_2 = 1$	0.62	0.668	0.681	0.675
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0.91	0.392	0.17	0.172
	$P_1 \& P_2 = -0.5$	0.906	0.83	0.808	0.74
	$P_1 \& P_2 = 0$	0.902	0.974	0.996	1
	$P_1 \& P_2 = 0.5$	0.914	0.782	0.744	0.772
	$P_1 \& P_2 = 1$	0.906	0.452	0.22	0.18

[a] 0,5,15,25 are the number of noise variables

Table 13: Sample Size Experiment Result In General Case

Criteria	RDD Model	Size	Simple	Medium	Complex
RMSE	**Linear Model**	1000	0.55	1.68	2.359
		2000	0.514	1.519	2.235
		3000	0.494	1.518	2.143
		4000	0.475	1.484	2.147
	Polynomial Model	1000	0.554	0.944	1.226
		2000	0.515	0.714	1.075
		3000	0.494	0.754	0.961
		4000	0.475	0.68	0.996
	PMWIT	1000	0.718	1.139	1.169
		2000	0.662	0.872	0.962
		3000	0.634	0.9	0.818
		4000	0.615	0.83	0.844
Mean Standard Error	**Linear Model**	1000	0.515	1.351	1.845
		2000	0.485	1.296	1.794
		3000	0.472	1.274	1.772
		4000	0.463	1.261	1.753
	Polynomial Model	1000	0.509	0.603	0.607
		2000	0.48	0.537	0.562
		3000	0.467	0.518	0.541
		4000	0.458	0.504	0.53
	PMWIT	1000	0.641	0.752	0.734
		2000	0.607	0.673	0.682
		3000	0.59	0.652	0.657
		4000	0.581	0.635	0.645
95% Confidence Interval Coverage	**Linear Model**	1000	0.905	0.87	0.861
		2000	0.907	0.883	0.864
		3000	0.908	0.884	0.869
		4000	0.912	0.883	0.871
	Polynomial Model	1000	0.891	0.877	0.751
		2000	0.896	0.898	0.716
		3000	0.898	0.897	0.707
		4000	0.905	0.901	0.697
	PMWIT	1000	0.886	0.863	0.871
		2000	0.892	0.887	0.881
		3000	0.894	0.887	0.89
		4000	0.9	0.891	0.894

Table 14: Table of Sample Size Test For Linear Model

Criteria/N Size	**Covariates**	**1000**	**2000**	**3000**	**4000**
RMSE	$P_1 \& P_2 = -1$	0.506	0.455	0.468	0.466
	$P_1 \& P_2 = -0.5$	0.485	0.462	0.409	0.438
	$P_1 \& P_2 = 0$	0.477	0.462	0.435	0.457
	$P_1 \& P_2 = 0.5$	0.481	0.474	0.459	0.414
	$P_1 \& P_2 = 1$	0.503	0.487	0.465	0.458
Mean Standard Error	$P_1 \& P_2 = -1$	0.492	0.479	0.464	0.445
	$P_1 \& P_2 = -0.5$	0.484	0.466	0.457	0.444
	$P_1 \& P_2 = 0$	0.458	0.46	0.45	0.439
	$P_1 \& P_2 = 0.5$	0.471	0.451	0.443	0.443
	$P_1 \& P_2 = 1$	0.504	0.471	0.470	0.436
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0.91	0.928	0.918	0.93
	$P_1 \& P_2 = -0.5$	0.934	0.92	0.938	0.918
	$P_1 \& P_2 = 0$	0.896	0.922	0.924	0.896
	$P_1 \& P_2 = 0.5$	0.898	0.904	0.898	0.942
	$P_1 \& P_2 = 1$	0.912	0.918	0.93	0.904

Table 15: Table of Sample Size Test For Polynomial Model

Criteria/N Size	**Covariates**	**1000**	**2000**	**3000**	**4000**
RMSE	$P_1 \& P_2 = -1$	0.507	0.463	0.477	0.463
	$P_1 \& P_2 = -0.5$	0.486	0.456	0.41	0.435
	$P_1 \& P_2 = 0$	0.471	0.459	0.425	0.454
	$P_1 \& P_2 = 0.5$	0.485	0.482	0.467	0.41
	$P_1 \& P_2 = 1$	0.509	0.488	0.465	0.453
Mean Standard Error	$P_1 \& P_2 = -1$	0.505	0.461	0.451	0.45
	$P_1 \& P_2 = -0.5$	0.477	0.459	0.446	0.439
	$P_1 \& P_2 = 0$	0.454	0.442	0.429	0.438
	$P_1 \& P_2 = 0.5$	0.56	0.574	0.578	0.584
	$P_1 \& P_2 = 1$	0.493	0.464	0.457	0.453
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0.9	0.912	0.906	0.914
	$P_1 \& P_2 = -0.5$	0.93	0.926	0.932	0.908
	$P_1 \& P_2 = 0$	0.898	0.914	0.922	0.908
	$P_1 \& P_2 = 0.5$	0.902	0.892	0.898	0.91
	$P_1 \& P_2 = 1$	0.89	0.906	0.906	0.916

Table 16: Table of Sample Size Test For PMWIT

Criteria/N Size	**Covariates**	**1000**	**2000**	**3000**	**4000**
RMSE	$P_1 \& P_2 = -1$	0.7	0.61	0.61	0.584
	$P_1 \& P_2 = -0.5$	0.604	0.592	0.547	0.568
	$P_1 \& P_2 = 0$	0.602	0.615	0.558	0.574
	$P_1 \& P_2 = 0.5$	0.618	0.581	0.605	0.539
	$P_1 \& P_2 = 1$	0.666	0.627	0.611	0.592
Mean Standard Error	$P_1 \& P_2 = -1$	0.634	0.594	0.588	0.566
	$P_1 \& P_2 = -0.5$	0.598	0.573	0.563	0.557
	$P_1 \& P_2 = 0$	0.571	0.571	0.552	0.558
	$P_1 \& P_2 = 0.5$	0.458	0.441	0.439	0.432
	$P_1 \& P_2 = 1$	0.607	0.582	0.582	0.566
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0.872	0.912	0.9	0.908
	$P_1 \& P_2 = -0.5$	0.906	0.91	0.922	0.912
	$P_1 \& P_2 = 0$	0.902	0.902	0.918	0.89
	$P_1 \& P_2 = 0.5$	0.9	0.914	0.882	0.91
	$P_1 \& P_2 = 1$	0.874	0.898	0.884	0.9

Table 17: Table of Sample Size Test For Polynomial Model In Pruning Forest

Criteria/N Size	**Covariates**	**1000**	**2000**	**3000**	**4000**
RMSE	$P_1 \& P_2 = -1$	2	1.03	0.633	0.416
	$P_1 \& P_2 = -0.5$	0.998	0.142	0.166	0.244
	$P_1 \& P_2 = 0$	0.134	0.22	0.2441	0.204
	$P_1 \& P_2 = 0.5$	1	0.144	0.175	0.242
	$P_1 \& P_2 = 1$	2.011	1.04	0.63	0.427
Mean Standard Error	$P_1 \& P_2 = -1$	0.134	0.131	0.118	0.113
	$P_1 \& P_2 = -0.5$	0.133	0.129	0.156	0.187
	$P_1 \& P_2 = 0$	0.134	0.248	0.252	0.223
	$P_1 \& P_2 = 0.5$	0.133	0.128	0.16	0.182
	$P_1 \& P_2 = 1$	0.134	0.13	0.118	0.113
Confidence Interval Coverage	$P_1 \& P_2 = -1$	0	0	0	0.06
	$P_1 \& P_2 = -0.5$	0	0.898	0.876	0.794
	$P_1 \& P_2 = 0$	0.956	0.948	0.936	0.938
	$P_1 \& P_2 = 0.5$	0	0.902	0.876	0.812
	$P_1 \& P_2 = 1$	0	0	0	0.044

Table 18: Second Stage Monte Carlo Simulation Result_Honest Fraction = 0.2

Criteria	**Model**	**25% Quantile**	**Median**	**75% Quantile**
RMSE	Linear	1.47	1.07	1.78
	Polynomial	1.69	1.2	1.84
	PMWIT	31.14	29.96	31.72
Mean Standard Error	Linear	0.835	0.998	0.974
	Polynomial	1.01	1.2	1.17
	PMWIT	27.53	40.66	39.73
Confidence Interval Coverage	Linear	0.61	0.91	0.598
	Polynomial	0.674	0.942	0.666
	PMWIT	0.926	0.98	0.974

Table 19: Second Stage Monte Carlo Simulation Result_Honest Fraction = 0.3

Criteria	**Model**	**25% Quantile**	**Median**	**75% Quantile**
RMSE	Linear	1.56	1.13	1.76
	Polynomial	1.77	1.29	1.88
	PMWIT	32.53	33.89	33.82
Mean Standard Error	Linear	1.02	1.35	1.26
	Polynomial	1.19	1.64	1.49
	PMWIT	35.12	59.68	52.79
Confidence Interval Coverage	Linear	0.706	0.976	0.784
	Polynomial	0.738	0.966	0.808
	PMWIT	0.952	0.986	0.982

Table 20: Auxiliary Covariates Description For extension case

Variable	**Description**
age	Age (years)

Table 20 Auxiliary Covariates Description

Variable	Description
sex	Female
raceblack	Black
raceother	Other
edu	Education (years)
income1	Income $11–$25k
income2	Income $25–$50k
income3	Income $>$ $50k
ins_care	Medicare
ins_paid	Private & Medicare
ins_caid	Medicaid
ins_no	No Insurance
ins_careaid	Medicare & Medicaid
cat1_copd	COPD
cat1_mosefsep	MOSF w/Sepsis
cat1_mosfmal	MOSF w/Malignancy
cat1_chf	CHF
cat1_coma	Coma
cat1_cir	Cirrhosis
cat1_lung	Lung Cancer
cat2_colon	Colon Cancer
cat2_mosfsep	MOSF w/Sepsis
cat2_coma	Coma
cat2_mosfmal	MOSF w/Malignancy
cat2_lung	Lung Cancer
cat2_cir	Cirrhosis
cat2_colon	Colon Cancer
resp	Respiratory diagnosis
card	Cardiovascular diagnosis
neuro	Neurological diagnosis
gastr	Gastrointestinal diagnosis

Table 20 Auxiliary Covariates Description

Variable	Description
renal	Renal diagnosis
meta	Metabolic diagnosis
hema	Hematological diagnosis
seps	Sepsis diagnosis
trauma	Trauma diagnosis
ortho	Orthopedic diagnosis
das2d3pc	DASI — Duke Activity Status Index
dnr1	Do Not Resuscitate status on day 1
ca_yes	Cancer — localized
ca_meta	Cancer — metastatic
surv2m1	Estimate of prob. of surviving 2 months
apache	APACHE score
glasgow	Glasgow coma score
wt1	Weight
temp1	Temperature
meanbp1	Mean Blood Pressure
resp1	Respiratory Rate
hr1	Heart Rate
paco2	PaCO2
ph	PH
wblc	WBC
hemal	Hematocrit
sodl	Sodium
potl	Potassium
creal	Creatinine
bilil	Bilirubin
albl	Albumin
cardiohx	Cardiovascular symptoms
chfx	Congestive Heart Failure
demhx	Dementia, stroke or cerebral infarct, Parkinson's disease

Table 20 Auxiliary Covariates Description

Variable	Description
psychhx	Psychiatric history, active psychosis or severe depression
chrpulhx	Chronic pulmonary disease, severe pulmonary disease
renalhx	Chronic renal disease, chronic hemodialysis or peritoneal dialysis
livrhx	Cirrhosis, hepatic failure
giblehdx	Upper GI bleeding
malighx	Solid tumor,metastatic disease, chronic leukemia/myeloma, acute leukemia, lymphoma
immunhx	Immunosuppression, organ transplant, HIV, Diabetes Mellitus, Connective Tissue Disease
transhx	Transfer ($>$ 24 hours) from another hospital
malhx	Definite myocardial infarction
wt0	weight = 0

Table 21: Comparison of generated and true data in extension simulation

Variable/RHC	**0**		**1**	
	Fake	**Real**	**Fake**	**Real**
Dth30	0.24 (0.37)	0.31 (0.46)	0.27 (0.39)	0.38 (0.49)
Age	56.39 (16.79)	61.76 (17.29)	55.53 (16.22)	60.75 (15.63)
Sex (Female)	0.46 (0.5)	0.46 (0.5)	0.41 (0.49)	0.41 (0.49)
Race (Black)	0.22 (0.41)	0.16 (0.37)	0.21 (0.4)	0.15 (0.36)
Race (Other)	0.08 (0.27)	0.06 (0.24)	0.08 (0.28)	0.07 (0.25)
Education	11.23 (3.04)	11.57 (3.13)	11.55 (3.05)	11.86 (3.16)
Income ($11-25k)	0.21 (0.41)	0.20 (0.4)	0.20 (0.4)	0.21 (0.41)
Income ($25-50k)	0.14 (0.34)	0.14 (0.35)	0.18 (0.38)	0.18 (0.38)
Income (>50k)	0.08 (0.27)	0.07 (0.26)	0.10 (0.30)	0.09 (0.28)
Medicaid	0.16 (0.37)	0.13 (0.33)	0.13 (0.33)	0.09 (0.28)
Medicare	0.28 (0.45)	0.27 (0.44)	0.25 (0.43)	0.23 (0.42)

Continued on next page

Table 21 continued from previous page

Variable	0		1	
	Fake	**Real**	**Fake**	**Real**
Medicare & Medicaid	0.10 (0.30)	0.07 (0.26)	0.06 (0.24)	0.06 (0.23)
No Insurance	0.08 (0.28)	0.05 (0.22)	0.09 (0.29)	0.06 (0.24)
Private & Medicare	0.22 (0.41)	0.21 (0.41)	0.23 (0.42)	0.22 (0.42)
COPD	0.13 (0.34)	0.11 (0.32)	0.03 (0.17)	0.03 (0.16)
MOSF w/Sepsis	0.18 (0.38)	0.15 (0.36)	0.37 (0.48)	0.32 (0.47)
MOSF w/Malignancy	0.06 (0.24)	0.07 (0.25)	0.09 (0.28)	0.07 (0.26)
CHF	0.08 (0.27)	0.07 (0.25)	0.10 (0.31)	0.10 (0.29)
Coma	0.11 (0.32)	0.10 (0.29)	0.05 (0.21)	0.04 (0.20)
Cirrhosis	0.06 (0.23)	0.05 (0.22)	0.04 (0.19)	0.02 (0.15)
Lung Cancer	0.01 (0.11)	0.01 (0.10)	0.01 (0.07)	0.00 (0.05)
Colon Cancer	0.00 (0.04)	0.00 (0.04)	0.00 (0.04)	0.00 (0.02)
MOSF w/Sepsis (Cat2)	0.13 (0.33)	0.11 (0.32)	0.19 (0.39)	0.19 (0.39)
Coma (Cat2)	0.02 (0.14)	0.02 (0.14)	0.02 (0.14)	0.01 (0.10)
MOSF w/Malignancy (Cat2)	0.07 (0.25)	0.05 (0.21)	0.03 (0.17)	0.03 (0.16)
Lung Cancer (Cat2)	0.00 (0.07)	0.00 (0.06)	0.00 (0.05)	0.00 (0.03)
Colon Cancer (Cat2)	0.00 (0.00)	0.00 (0.02)	0.00 (0.00)	0.00 (0.02)
Cirrhosis (Cat2)	0.01 (0.09)	0.01 (0.09)	0.01 (0.09)	0.01 (0.07)
Resp	0.42 (0.49)	0.42 (0.49)	0.30 (0.46)	0.29 (0.45)
Card	0.28 (0.45)	0.28 (0.45)	0.43 (0.50)	0.42 (0.49)
Neuro	0.18 (0.38)	0.16 (0.37)	0.06 (0.23)	0.05 (0.23)
Gastr	0.17 (0.38)	0.15 (0.35)	0.23 (0.42)	0.19 (0.39)
Renal	0.06 (0.24)	0.04 (0.20)	0.08 (0.28)	0.07 (0.25)
Meta	0.06 (0.24)	0.05 (0.21)	0.06 (0.24)	0.04 (0.20)
Hema	0.07 (0.26)	0.07 (0.25)	0.07 (0.25)	0.05 (0.22)
Seps	0.19 (0.39)	0.15 (0.35)	0.26 (0.44)	0.24 (0.42)
Trauma	0.02 (0.14)	0.01 (0.07)	0.02 (0.13)	0.02 (0.12)
Ortho	0.00 (0.05)	0.00 (0.03)	0.00 (0.04)	0.00 (0.04)
das2d3pc	18.15 (5.52)	20.37 (5.48)	18.42 (5.22)	20.70 (5.03)
Dnr1	0.17 (0.37)	0.14 (0.35)	0.09 (0.29)	0.07 (0.26)

Continued on next page

Table 21 continued from previous page

Variable	0		1	
	Fake	**Real**	**Fake**	**Real**
Ca (Yes)	0.19 (0.40)	0.18 (0.38)	0.14 (0.35)	0.15 (0.36)
Ca (No)	0.75 (0.43)	0.75 (0.43)	0.80 (0.40)	0.79 (0.41)
Surv2md1	0.63 (0.19)	0.61 (0.19)	0.61 (0.19)	0.57 (0.20)
Aps1	49.78 (18.95)	50.93 (18.81)	59.40 (20.19)	60.74 (20.27)
Scoma1	21.48 (32.30)	22.25 (31.37)	16.52 (25.85)	18.97 (28.26)
Wtkilo1	70.29 (28.87)	65.04 (29.50)	76.14 (27.69)	72.36 (27.73)
Temp1	38.17 (1.75)	37.63 (1.74)	38.16 (1.77)	37.59 (1.83)
Meanbp1	83.62 (37.18)	84.87 (38.87)	67.22 (34.75)	68.20 (34.24)
Resp1	28.98 (13.96)	28.98 (13.95)	26.66 (14.15)	26.65 (14.17)
Hrt1	110.06 (39.82)	112.87 (40.94)	117.72 (41.45)	118.93 (41.47)
Pafi1	253.83 (106.35)	240.63 (116.66)	211.55 (107.55)	192.43 (105.54)
Paco21	42.23 (13.65)	39.95 (14.24)	39.25 (11.79)	36.79 (10.97)
Ph1	7.39 (0.11)	7.39 (0.11)	7.38 (0.11)	7.38 (0.11)
Wblc1	15.16 (9.71)	15.26 (11.41)	16.45 (10.36)	16.27 (12.55)
Hema1	32.23 (8.61)	32.70 (8.79)	30.00 (7.90)	30.51 (7.42)
Sod1	139.41 (7.58)	137.04 (7.68)	138.71 (7.70)	136.33 (7.60)
Pot1	4.11 (0.98)	4.08 (1.04)	4.14 (1.04)	4.05 (1.01)
Crea1	2.11 (1.89)	1.92 (2.03)	2.65 (2.09)	2.47 (2.05)
Bili1	2.07 (3.35)	2.00 (4.43)	2.34 (3.40)	2.71 (5.33)
Alb1	3.15 (0.66)	3.16 (0.67)	2.96 (0.68)	2.98 (0.93)
Cardiohx	0.17 (0.38)	0.16 (0.37)	0.23 (0.42)	0.20 (0.40)
Chfhx	0.18 (0.38)	0.17 (0.37)	0.22 (0.41)	0.19 (0.40)
Dementhx	0.14 (0.35)	0.12 (0.32)	0.10 (0.30)	0.07 (0.25)
Psychhx	0.09 (0.28)	0.08 (0.27)	0.06 (0.24)	0.05 (0.21)
Chrpulhx	0.25 (0.43)	0.22 (0.41)	0.16 (0.37)	0.14 (0.35)
Renalhx	0.05 (0.21)	0.04 (0.20)	0.06 (0.23)	0.05 (0.21)
Liverhx	0.09 (0.28)	0.07 (0.26)	0.07 (0.25)	0.06 (0.24)
Gibledhx	0.05 (0.21)	0.04 (0.19)	0.04 (0.20)	0.02 (0.16)
Malighx	0.25 (0.43)	0.25 (0.43)	0.20 (0.40)	0.20 (0.40)

Continued on next page

Table 21 continued from previous page

Variable	**0**		**1**	
	Fake	**Real**	**Fake**	**Real**
Immunhx	0.32 (0.47)	0.26 (0.44)	0.35 (0.48)	0.29 (0.45)
Transhx	0.12 (0.33)	0.09 (0.29)	0.18 (0.39)	0.15 (0.36)
Amihx	0.04 (0.19)	0.03 (0.17)	0.04 (0.19)	0.04 (0.20)
Wt0	0.11 (0.31)	0.10 (0.30)	0.07 (0.26)	0.07 (0.25)

Table 22: Comparison of generated data and true data in second stage Monte-Carlo simulation

Treatment	0		1	
Source	**Fake**	**Real**	**Fake**	**Real**
Poverty Index	34.08(11.66)	33.29(12.04)	66.19(8.04)	66.05(5.58)
Mortality rate per 100,000	1.83(2.88)	2.23(5.83)	2.01(4.00)	2.42(4.47)
Total Population	41602(125258)	41521(123775)	17513(16206)	17566(16390)
Attending school%, age 14-17	85.37(8.94)	84.44(16.77)	80.24(8.98)	81.01(10.75)
Attending school%, age 5-34	0.53(0.05)	0.55(0.06)	0.56(0.05)	0.57(0.05)
High-school or more, age 25+	32.29(9.29)	34.80(9.56)	17.50(6.65)	19.68(4.96)
Population, age 14-17	2695(7041)	2690(6957)	1545(1338)	1552(1352)
Population, age 5-34	19443(55734)	19405(55133)	8958(8534)	9001(8636)
Population, age 25+	23310(74530)	23270(73791)	8649(7330)	8677(7405)
Urban population%	35.10(27.35)	31.00(26.93)	12.70(16.39)	13.71(18.04)
Black population %	8.12(13.48)	7.91(12.81)	35.42(25.54)	33.45(26.31)

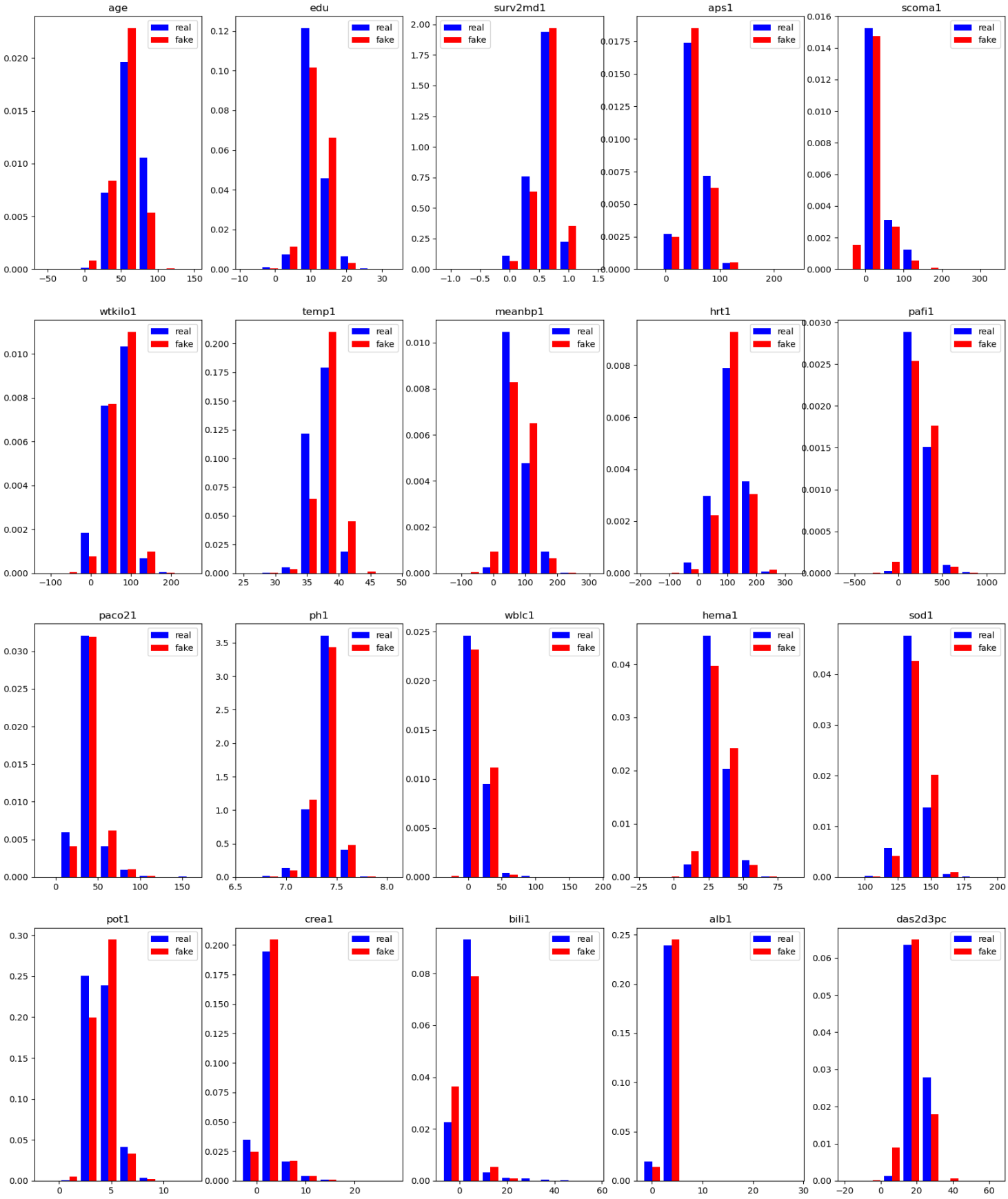

Figure 7: Continuous Variable Comparison

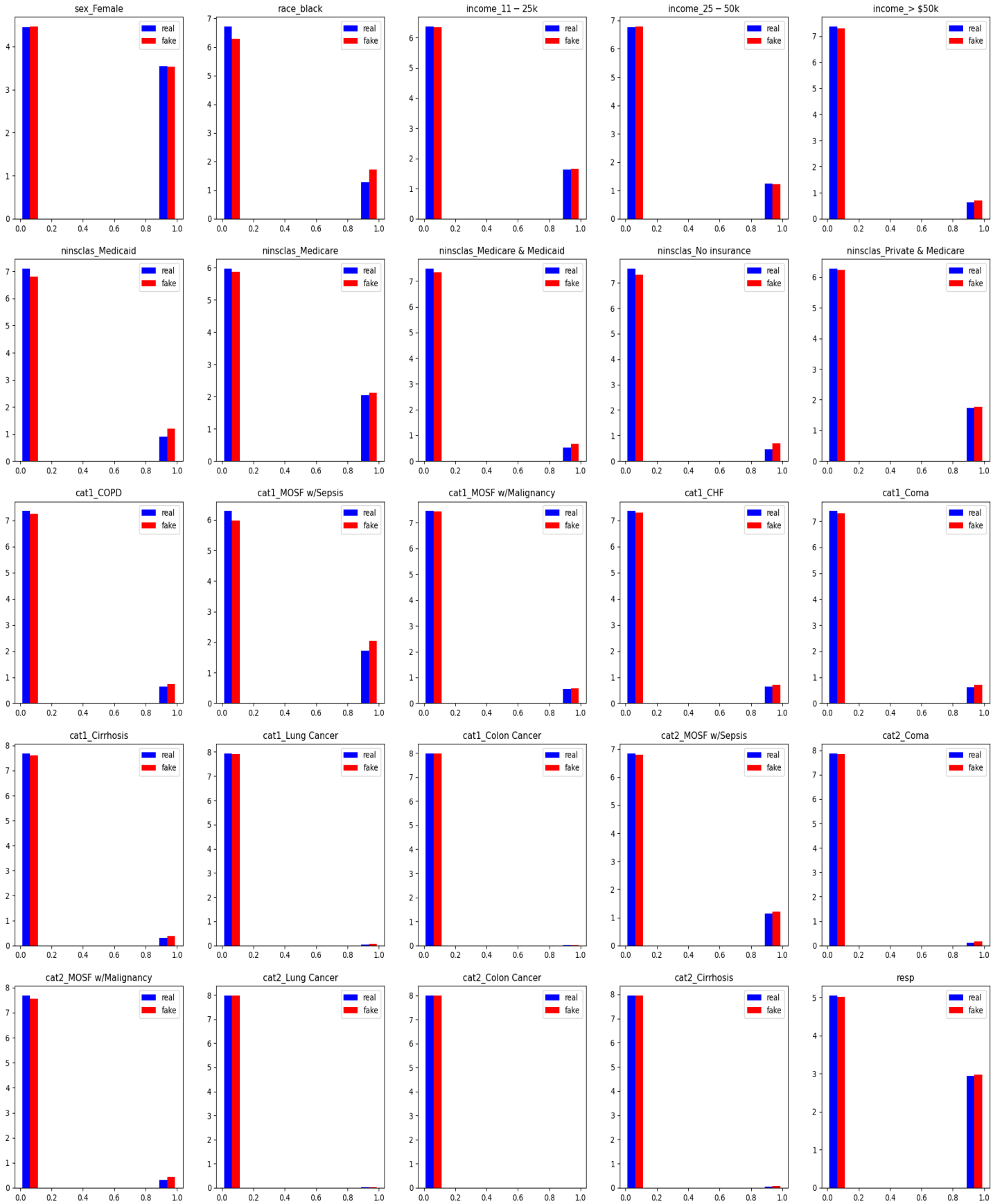

Figure 8: Categorical Variable Comparison-1

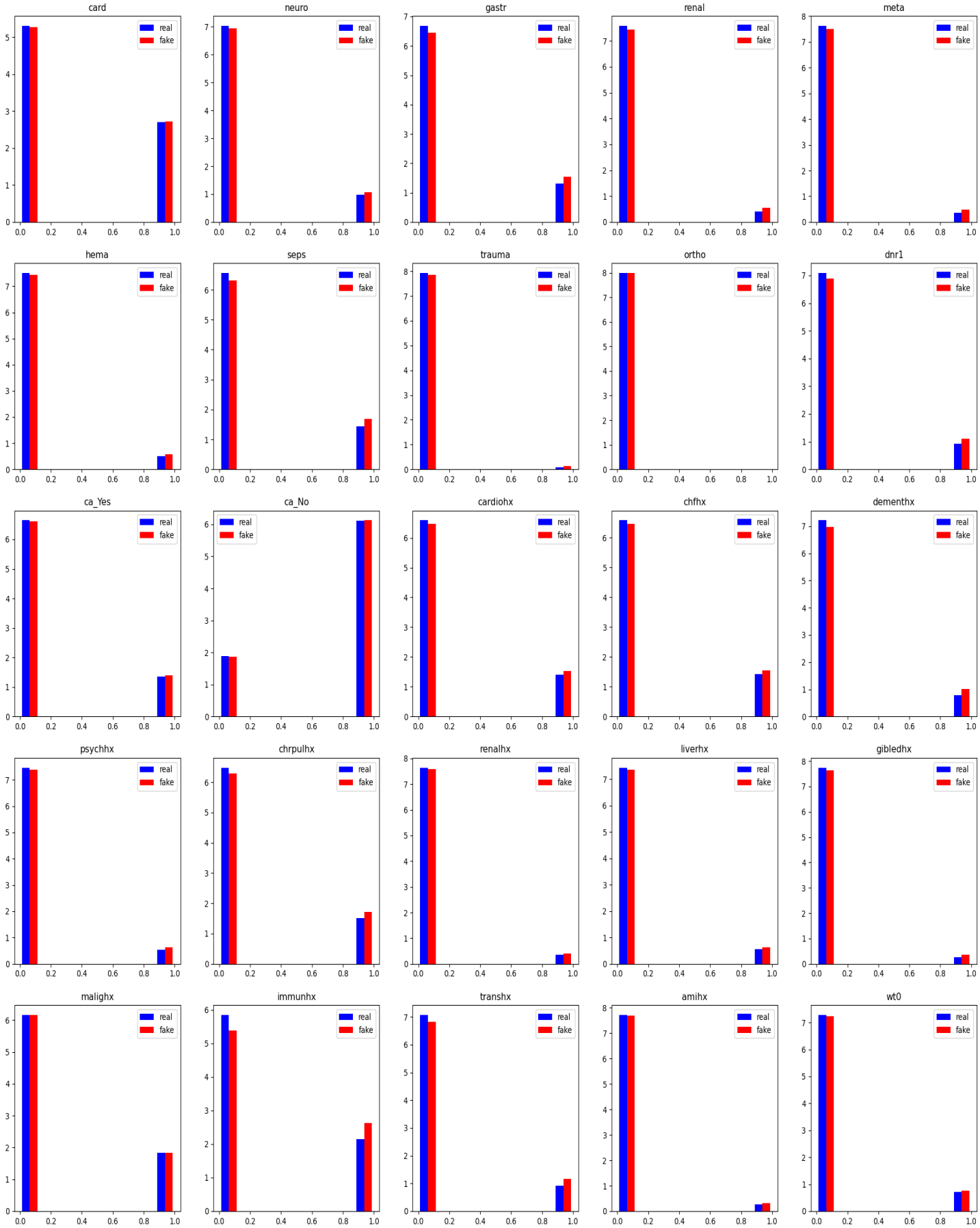

Figure 9: Categorical Variable Comparison-2

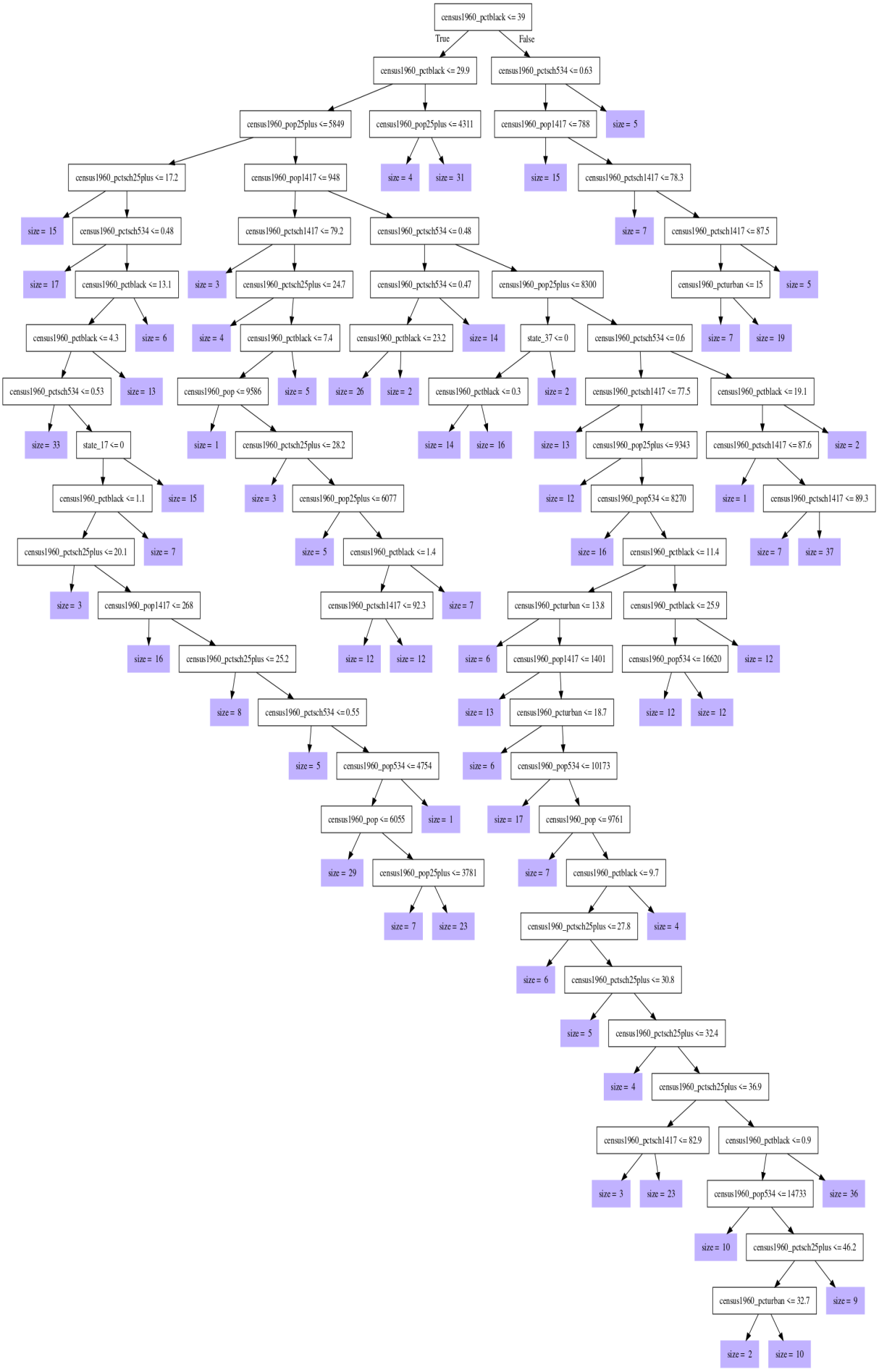

Figure 10: Poly$_2$5*Forest*

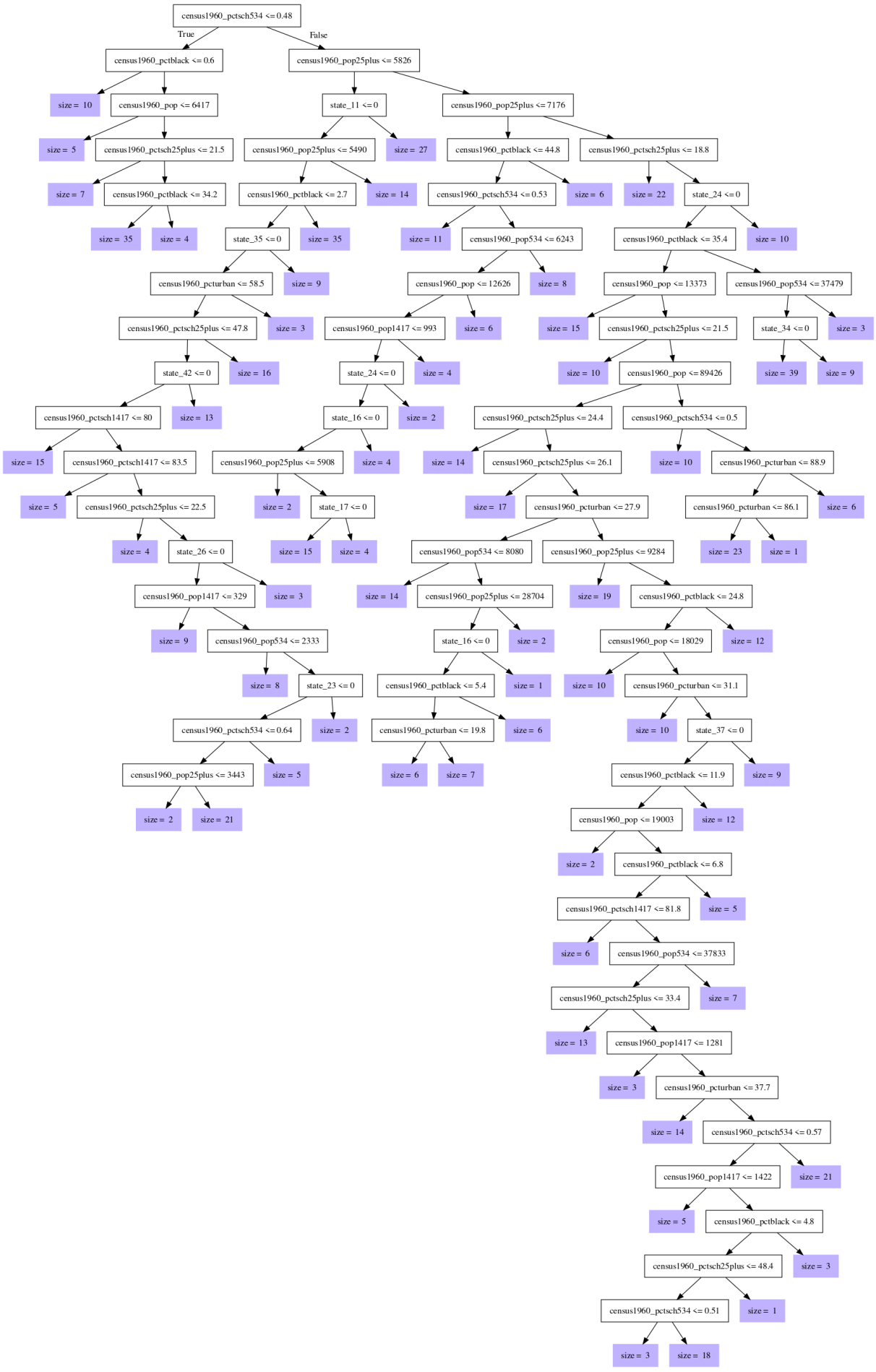

Figure 11: Linear_25 Forest

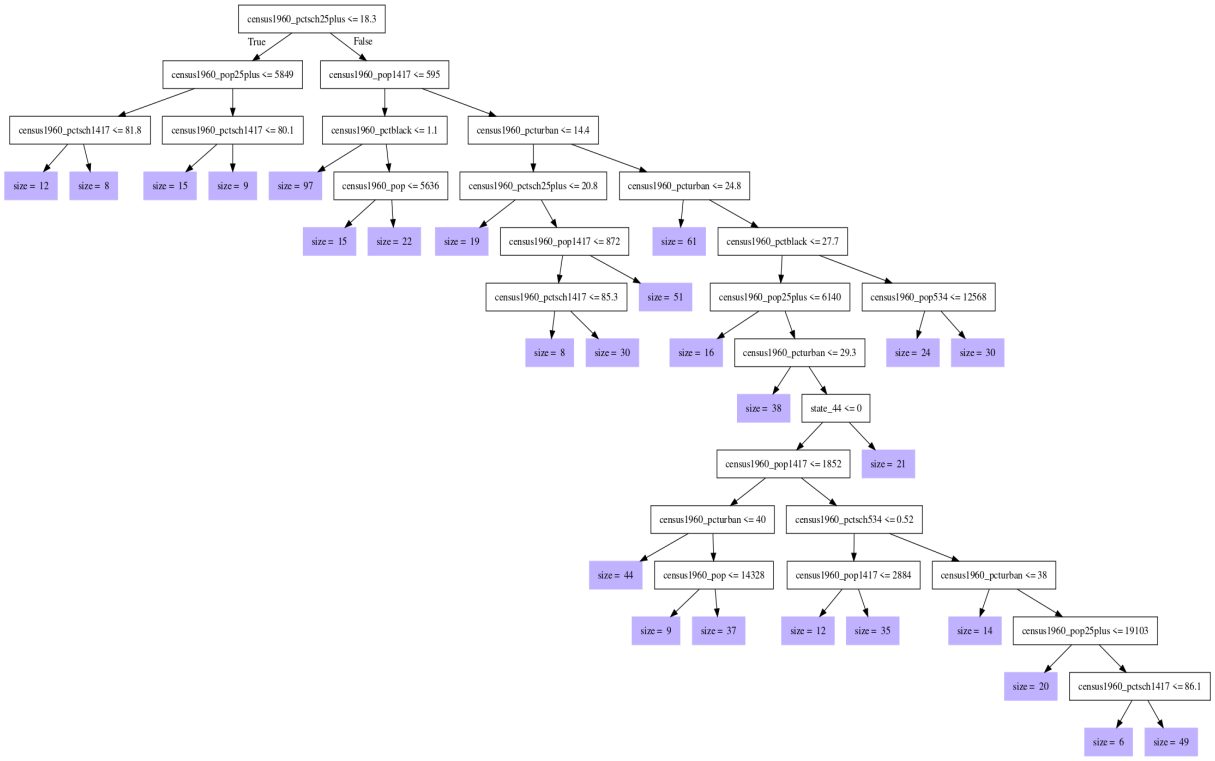

Figure 12: UR_poly

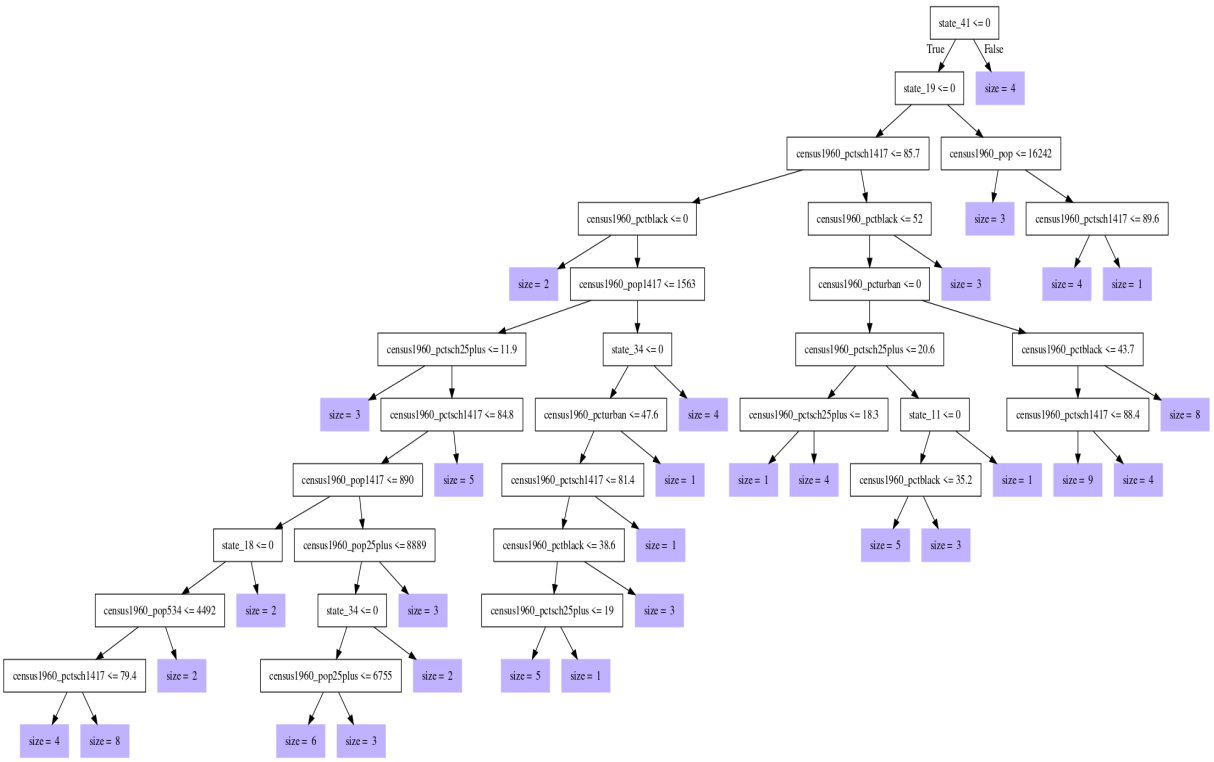

Figure 13: Bandwidth Poly

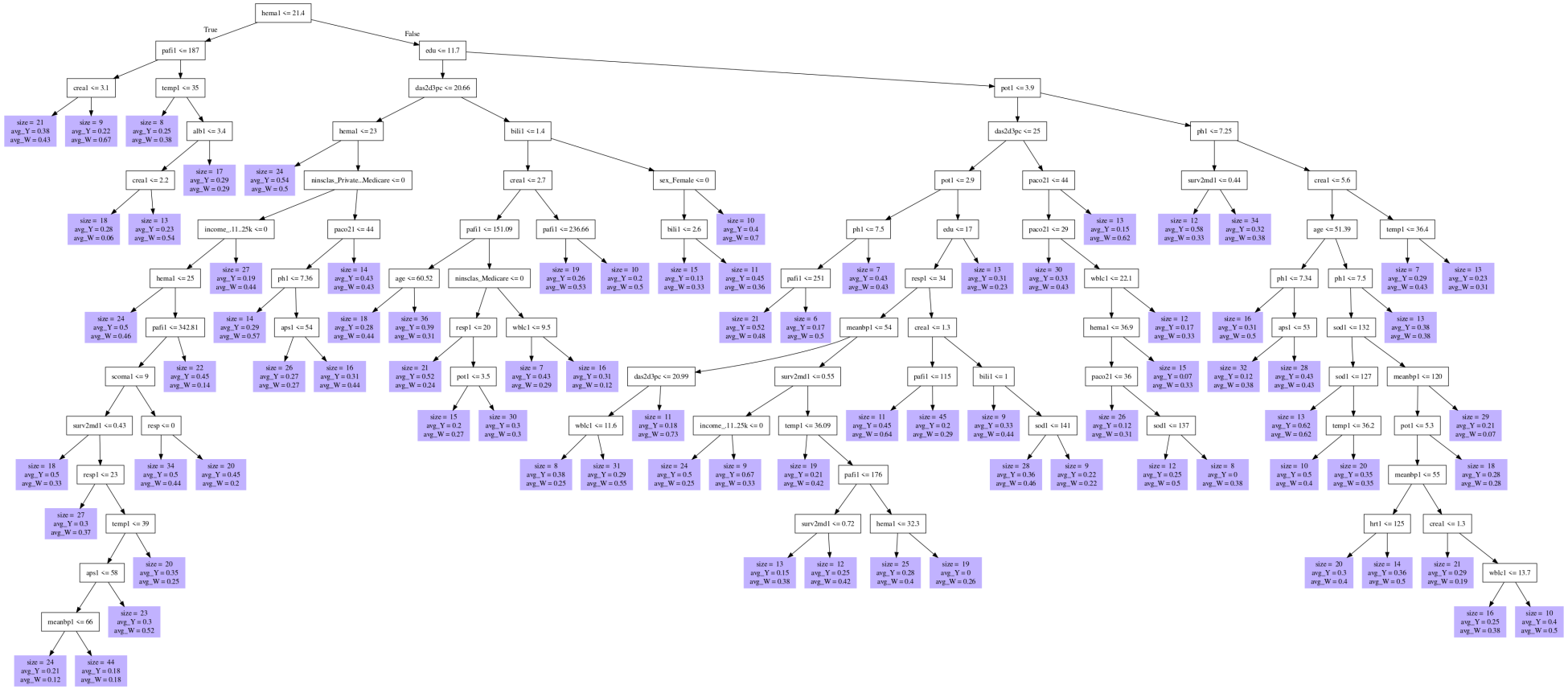

Figure 14: 4000 Tree Forest

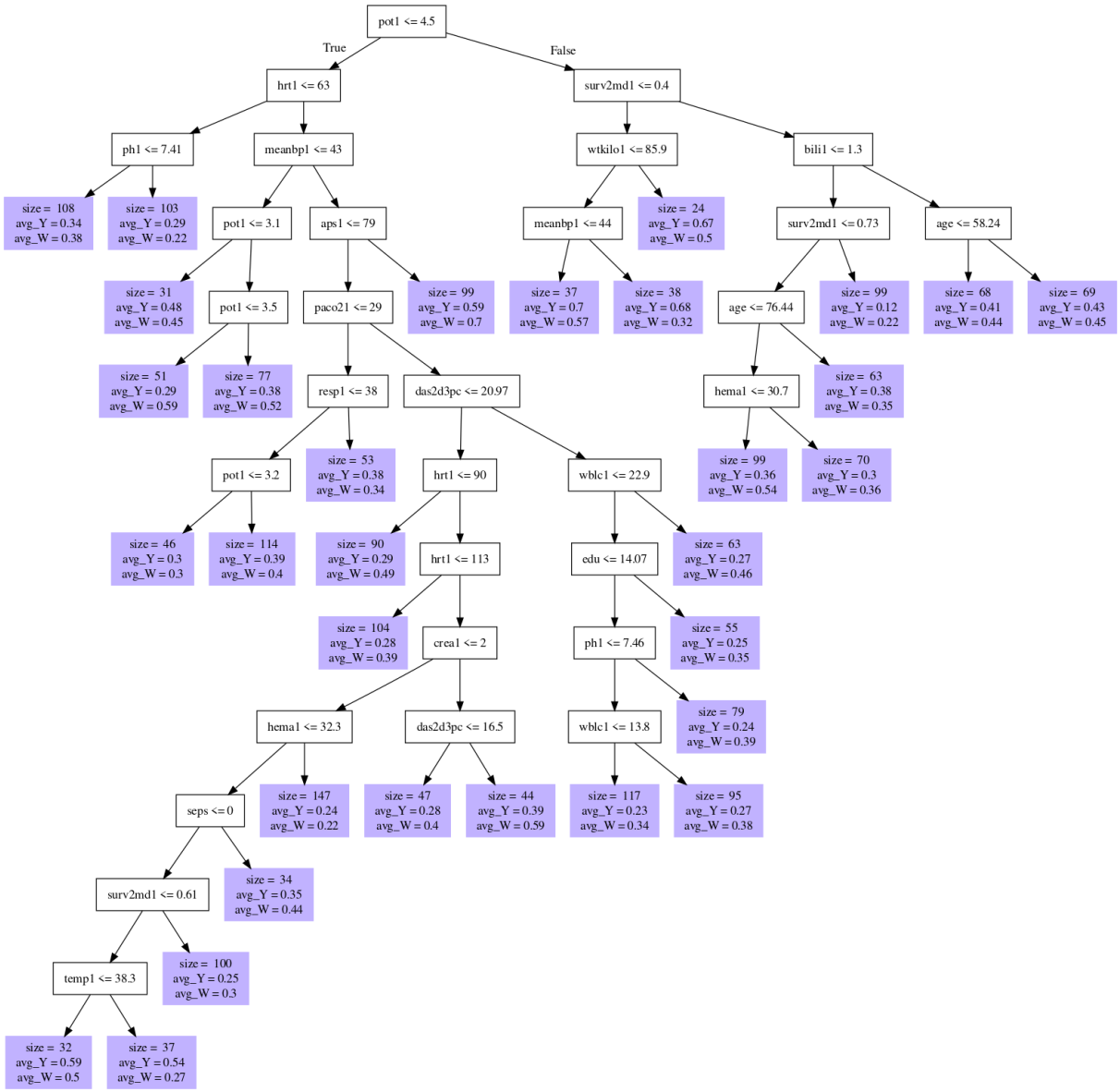

Figure 15: 0.2 Honest Forest

8.4 Extension Discussion

8.4.1 WGAN

Following will be an introduction about how the WGAN method works in the thesis. First, about the GAN model, as mentioned before, there are two important neural nets inside, which are the discriminator and generator. The generator can be written in the following form: $g(O_i;\theta_g)$, and normally O_i follows normal distribution or uniform distribution. θ_g is the weight in the neural nets to modify the distribution of given O_i to mimic the real-world data. Besides, the discriminator can be written as $d(Q_i;\theta_d)$, Q_i is the real-world data, and θ_d is again the weight of neural nets. It is important to note that in the hidden layer of the discriminator, there is no determined activation function, and it can be a non-linear activation function to let the model have a strong expressive capability. But for the output layer, usually it applied Sigmoid activation function, therefore regardless how big is value, it will turns into the value between 0 and 1 like probability. In other words, the discriminator would return a value like the probability to tell whether the input data is original or not. Combining these notations, we can have the following Jensen-Shannon divergence.

$$\min_{\theta_g}\max_{\theta_d} JS(\theta_d,\theta_g) = ln2 + \frac{1}{N_R}\sum_{i=1}^{N_R}\ln d(Q_i;\theta_d) + \frac{1}{N_F}\sum_{i=1}^{N_F}\ln[1-d(g(O_i;\theta_g);\theta_d)]$$

N_R and N_F stand for the number of real and fake samples. In this optimization question, people are trying to maximize the equation by finding the optimal θ_d, also to minimize the equation by searching the optimal θ_g. In the process, the gradient will be calculated iterately. However, the common issue with traditional GAN is its instability. It is because the discriminator may improve too fast, so it always returns the value close to one or zero for the output layer, and then it becomes challenging for the generator to have significant improvement. This is usually the case when the generator and the real generating process are two disjoint distribution Gulrajani et al. (2017). The discriminator can easily distinguish true from fake, and then the generator has a gradient vanish problem, which in the end, leads to mode collapse. To overcome this problem, Gulrajani et al. (2017) suggesting the WGAN model, which estimate the distance between two distribution. The main modification from WGAN is that in the output layer it applied linear activation, which is not limit the number between 0 and 1, but giving a score of result. But inside the neural nets, the hidden layer can still be other non-linear activation, instead of just using the linear activation, otherwise it will makes the model to simple(just linear combination), which is hard to mimic the real world data. The WGAN model can be written like following

Athey et alAthey et al. (2024).

$$\min_{\theta_g}\max_{\theta_d}\left\{\frac{1}{N_R}\sum_{i=1}^{N_R} f(Q_i;\theta_c)-\frac{1}{N_F}\sum_{i=1}^{N_F} f(g(O_i;\theta_g);\theta_c)\right\}$$

Beside changing the the output layer activation function, Gulrajani et al. (2017) did another step, weight clipping, to increase the stability of model. They restrict the function to Lipschitz being smaller or equal to one. In other words, they want to make sure the following inequality hold. $\left|f(\hat{Q_{i1}})-f(\hat{Q_{i2}})\right|\leq\left|\hat{Q_{i1}}-\hat{Q_{i2}}\right|$. Here $\hat{Q_{i1}}$ and $\hat{Q_{i2}}$ is a random linear combination between original data and generated data. To ensure that in the process of making generated data looks like original data, the gradient is stable.

In the end, this constrain would turns into a penalty term, adding in the optimization function.

$$\min_{\theta_g}\max_{\theta_d}\left\{\frac{1}{N_R}\sum_{i=1}^{N_R} f(Q_i;\theta_c)-\frac{1}{N_F}\sum_{i=1}^{N_F} f(g(O_i;\theta_g);\theta_c)+\lambda\frac{1}{m}\sum_{i=1}^{m}\left[\max\left(0,\left\|\nabla_{\hat{Q}_i} f(\hat{Q}_i;\theta_c)\right\|_2-1\right)\right]^2\right\} \tag{14}$$

Notice that, after obtaining the gradient from the above optimization question, we need to plug it into the adaptive moment estimation(ADAM), for further stabilize the result.

8.4.2 Extension: Second Stage Monte Carlo Simulation in a general case

In the extension part, a complex dataset will be applied. It has over 70 variables, including various continuous and categorical data. After generating the estimated data, the model equation1 defined in the beginning of the thesis will be applied and there will be a short simulation testing different properties about the GRF. Our data is from Connors et al. (1996). The research that Connors et al. (1996) was doing is to estimate the treatment effects of right heart catheterization surgery on the mortality rate of the patients. Surprisingly, their result shows that the right heart catheterization surgery may increase the mortality rate, which is counter-intuitive to our thoughts. Hirano and Imbens (2001) using a different treatment affects estimating methods to estimate the the effects of right heart catheterization and obtain a similar outcome. Before proceeding with the simulation, we can find the complete list of variables in the appendix. All the applied variables are consistent with the paper from Hirano and Imbens (2001). Next, three histograms are presented to illustrate the comparison of real data distribution and the generated data distribution from the neural network. 50 categorical covariates and 20 continuous covariates are plotted. These charts can again can be found in the appendix. From the graph, we can easily find that all the generated variables are quite similar to the distribution of real data. To

see the comparison further, we can see the table in the appendix, which lists all the mean and variance of variables and compare them with their true mean and deviation. Again, for some variables with small means and standard deviation, like "Lung Cancer", "Colon Cancer" and "Cirrhosis", the generated data can simulate them well. And for large mean and standard error variables like "pafi 1" ,which is the variable to classify severity of acute respiratory distress syndrome (ARDS), the mean of original data is 240 and standard deviation is 116 for control group and in the same group the generated data has 253 for mean and 106 for standard deviation. In conclusion, from this table we can knows that the WGAN can also perform well for generating a relatively large, complex, various variable dataset.

From The original data, it has a total of 5735 observations, and there are 2184 treated patients and 3551 control patients. In the generated data , there are, again in total 500,000 data points, and among them, 309143 observations are treated, and 190857 are controlled. Next, when doing the extension of the second stage of the Monte-Carlo simulation, I will again consider three "hypothetical" points, which are located at 25% quantile, median, and 75% quantile. The model from equation1 will be applied. In this dataset, only the treatment variables would be included in the local regression model. Based on the model, three additional setup will be tested. The first model would be applied is a regular forest with 2000 trees as a benchmark; next, 4000 tree forests would be applied. Besides, another setup will change the fraction of honest tree. The initial value is 0.5 of the sample would be taken for splitting, and another 0.5 would be taken for estimating. Now it would be 0.2 for splitting and 0.8 for estimating the result. The simulation would be iterated 500 times. The result from this three setup would be evaluated by the criteria we used before: RMSE, bias, standard error, and confidence interval.

In the table, we can see that the RMSE is almost the same for all the points. Only 75% quantile has a slightly better result. Also,the result fits the general concept of a normal random forest that usually increases the number of trees and will not have a significant impact on the bias, but it affects the variance. Here from the table, we can also see that 4000 tree forest has lower standard error than 2000 tree. Moreover, when it comes to honest fractions, now because the number of observations for estimation increases (the number for splitting decreases), therefore, it also reflects on the standard error, that it has a more consistent result, not necessarily better for performance but lower down the mean, standard error. In terms of bias, unlike our previous simulation, there is no clear trend regarding which model has an upward or downward bias; it also seems all the models have some outlier points. In the end, three models have quite similar results, and they will be applied again in the extension of the application section.

Table 23: Second Stage Monte Carlo Simulation Result

Criteria	Model	25% Quantile	Medium	75% Quantile
RMSE	2000 Tree	0.034	0.034	0.03
	Honest Fraction	0.029	0.03	0.027
	4000 Tree	0.034	0.034	0.03
Mean Standard Error	2000 Tree	0.07	0.05	0.054
	Honest Fraction	0.037	0.028	0.03
	4000 Tree	0.057	0.042	0.046
Confidence Interval Coverage	2000 Tree	0.996	0.982	0.996
	Honest Fraction	0.97	0.896	0.94
	4000 Tree	0.99	0.964	0.988

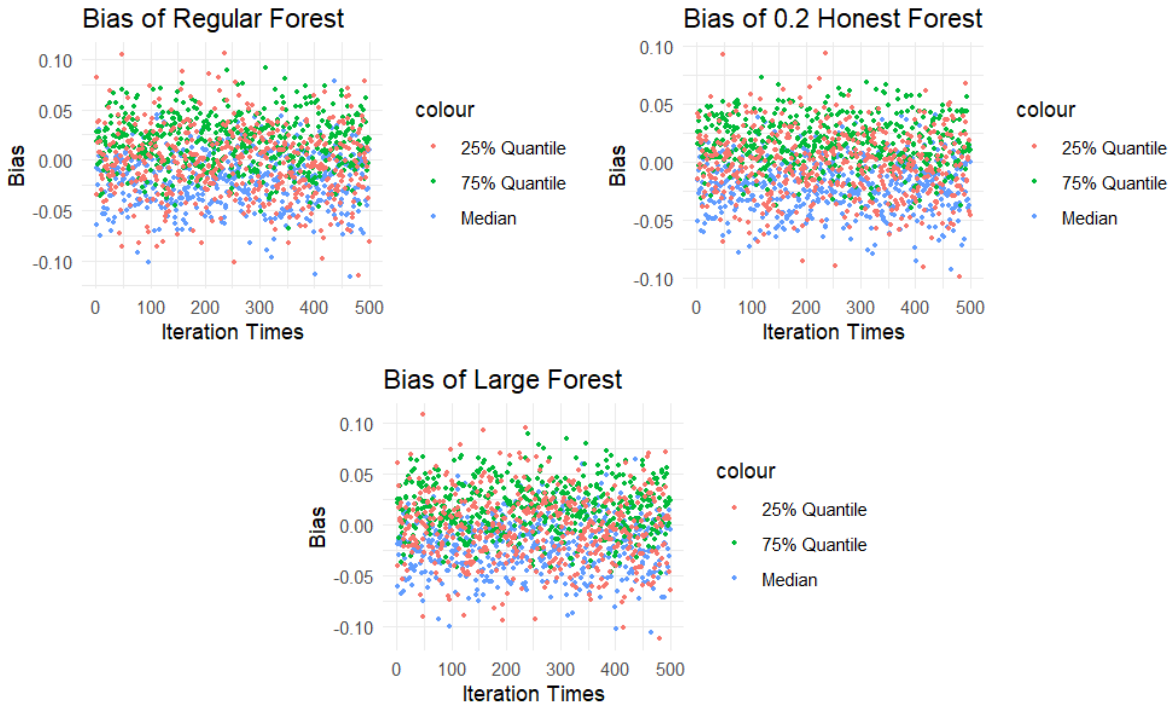

Figure 16: Estimation Bias

8.4.3 Extension of Application:Right Heart Catheterization

For the extension of the application, the data from Connors et al. (1996) would be applied again. In the real data, it has 5735 observations and among them 3551 are control and 2184 are treated. First, I applied the same processing method from the paper Hirano and Imbens (2001)

and McConnell and Lindner (2019) to process the original data from Vanderbilt Biostatistics Dataset[2]. Next, I again applied the three GRF models to do the estimation, which are 2000 tree causal forests, 4000 causal forests, and changing the honest fraction to 0.2 for splitting. The estimation result is shown the following table.

Table 24: RHC Estimation Result

Model	**Average Treatment Effect**	**Standard Error**
Regular Forest	4.476%	1.311%
Honest Fraction	4.317%	1.311%
Larger Forest	4.402%	1.309%
Benchmark	4.3%	1.3%

[1] Benchmark is from the paper McConnell and Lindner (2019) McConnell and Lindner (2019)

Note that now all the estimations are with the unbalanced restriction. Furthermore, as you can see, the estimation result is close to the result in McConnell and Lindner (2019). All the differences are about 0.1%, and the standard error estimation is the same as the paper. Now, based on this result, we will have further discussion about the heterogeneity of the estimation among these three models.

Please reference the tree from the 4000 tree forests and the 0.2 honest fraction trees from the appendix. In the thesis, we have introduced the restriction of unbalanced allocation several times, but we have not seen it in the result. Now with the causal forest package in R, we can not only observe the structure of the tree, but see the detail about the ratio of observation being treated in the leaf. We can clearly see that both of these two trees are with the restriction. In the end, it turns out that almost none of the leaf have the issue of disproportional distribution of the observations. Although the proportion of receiving the treatment is not always 0.5 but at least almost all of them were maintained at a certain level.

To check more details about one of the tree from 0.2 fraction forest. There are 0.2 proportions of data being used for partitioning and 0.8 for estimation. Other than these settings, all the tuning parameters are the same as the 2000 tree forest. Obviously, this adjustment is reflected in the graph. We can observe that the observations in the leaf is considerably more than the one

[2]Vanderbilt Biostatistics Dataset: https://hbiostat.org/data/

in 4000 tree model figure.

Next, we dive into the heterogeneity analysis. First, I plotted the heterogeneous treatment effects distribution again for both 2000 tree versus 4000 tree causal forest and 2000 tree versus 0.2 fraction honest tree causal forest. At the bottom of the page, we can see that the one on the left-hand side is a 2000-tree causal forest versus a 4000-tree causal forest. In terms of their distribution, it does not shows a discernible difference, the green color bar is the distribution of 2000 tree forest and blue one is 4000 tree forest. Besides, it is clear it shows a positive estimated heterogeneous treatment effect for most of the observations, which has the same point of view from Connors et al. (1996), Hirano and Imbens (2001) and (2019) McConnell and Lindner (2019) that the right heart catheterization would increase the mortality rate from ATE perspective. The figure on the right-hand side is the 2000 tree causal forest versus 0.2 fractions honest forest. Again, the green one is a 2000-tree forest, and the blue one is a 0.2 fraction of the honest forest. By changing the the fraction of honest trees, we can see that now, because there are more observations applied for estimation in each leaf instead of splitting, so by taking account of them in the estimation process, the graph becomes more centering than 2000 tree forest.

Next, to quantify the level of heterogeneous effects in each model estimation, the rank-weighted average treatment effects proposed from the paper by Yadlowsky et al(2021) **?** will be applied. To give a brief introduction, there are some notation need to be defined. First, we have a function $S(P_i)$. Here, S is a function to rank which observation has strong heterogeneous treatment effects(HTE), and P_i is again the auxiliary covariates. Now, we can calculate the targeting operator characteristic(TOC) by following formula.

$$\text{TOC}(q) = E[Y_i(1) - Y_i(0) \mid S(X_i) \geq F^{-1}_{S(X_i)}(1-q)] - E[Y_i(1) - Y_i(0)], \tag{15}$$

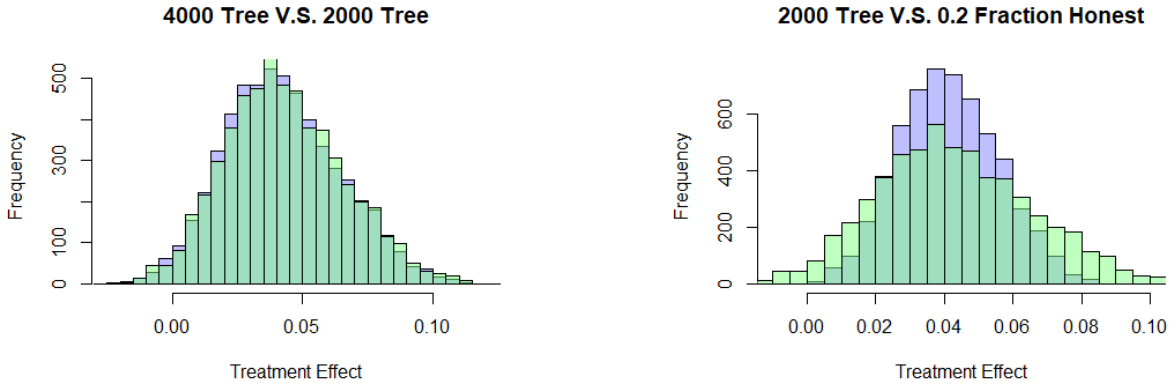

Figure 17: Heterogeneous Treatment

F is the distribution of ranking function S. To understand the formula intuitively, we can see when q is equal to almost; it is calculating that for those observations who have high ranking treatment effects, how much average treatment effects they have and compare it with the general average treatment effects. If it shows a strong difference, it means that the estimation result is highly heterogeneous. When q starts to grow, the difference should also decrease, and until q is equal to one, the difference would be zero.

In our estimation, the equation above was changed to the following form.

$$\mathrm{TOC}(q) = E\left[Y_i^{\text{test}}(1) - Y_i^{\text{test}}(0) \mid \hat{T}_i^{\text{train}}(X_i^{\text{test}}) \geq \hat{F}^{-1}(1-q)\right] - E\left[Y_i^{\text{test}}(1) - Y_i^{\text{test}}(0)\right] \quad (16)$$

First, I sample my data to train and test data, and both train and test data have the same proportion of treated and control observations. Next, I used the train data to train the first 2000 tree causal forest, which is $\hat{T}_i^{\text{train}}$ in the above equation and take it as the ranking function mentioned above. Next, I train the second causal forest from test data. When I was calculating the TOC, I put the test data into the first trained causal forest model to obtain the HTE and then take it as the rank to rank all the observations. When q = 0, I calculate the difference between the highest rank observations' average treatment effects from the second-trained causal forest and the general average treatment effects also from the second-trained causal forest. When q increase, the first expectation term includes more observations and again when q = 1 the first expectation term would be the same as second term.

From the plot below, we can see the process of different levels of heterogeneity when q increases from zero to one. The solid line in the graph is the TOC estimation, and the dash line is the confidence interval. From the graph, we can see that before q = 0.2, the average treatment

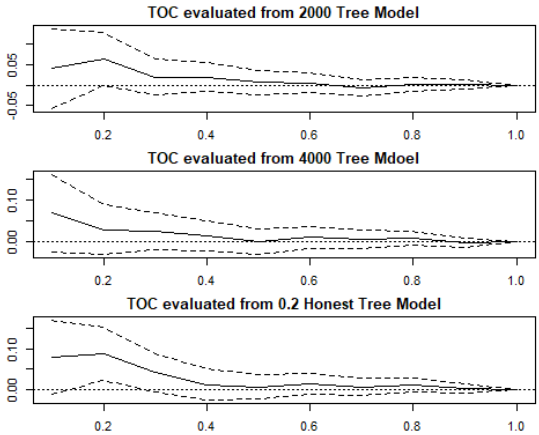

Figure 18: Heterogeneous Treatment Effects, xlab = q, ylab = Difference

effects from those high-rank observations show a difference from the general average treatment effects, and it gradually decreases. When q = 0.5, we can see two expectation terms almost merged together and this is the case until q = 1. Following the area under the TOC line would be calculated like the following.

$$\text{RATE} = \int_0^1 \text{TOC}(q)\, dq \tag{17}$$

We take the integral from q=0 to q=1 to numerically describe the HTE in each model. Following table are the result of RATE estimation for three models.

Table 25: RATE Estimation Result

Model	**RATE**	**Standard Error**
Regular Forest	2.048%	1.567%
Honest Fraction	2.118%	1.561%
Larger Forest	2.559%	1.586%

[1] The estimation methed is "AUTOC"

Notice that from here, we can also calculate the p-value of the estimation and see if the RATE is significantly different from zero. From the table here, we can derive that the p-value for the three models is not significant. However, the results for each time estimation can be different, and my research did not consistently show a significant result(or insignificant result). On top of that, I also found for those round which has significant HTE(p-value $<$0.1), if I do not re-sampling them again but use the same train and test data to construct another two forests, one for ranking and one for evaluation. Then, the result would usually be significant again. Therefore, it shows that there are some important observations in the sample that can make the HTE estimation stronger, and the prerequisite to significantly observe them from RATE is that those observations need to be balanced and allocated to the first and second causal forests. From the application, we can observe that, besides the correct specification of the model for HTE estimation, enough of data points to show the HTE may also be important. Otherwise, it may be the case that when using the entire dataset, it has a significant heterogeneity level, but in the RATE, the RATE disappear.

Statement of authorship:

I hereby confirm that the work presented has been performed and interpreted solely by myself except for where I explicitly identified the contrary. I assure that this work has not been presented in any other form for the fulfillment of any other degree or qualification. Ideas taken from other works in letter and in spirit are identified in every single case.

Bonn, 29.08.2024 _____

Yu-Hsin Chen