Let 1,000 Flowers Bloom (or Wilt): Heterogeneity in National Market-Level Charter School Effects – ML Application

Feng Chen, Douglas N. Harris, and Nicholas Lacoste

November 13, 2024

1 Exploring Heterogeneity Nonparametrically via Generalized Random Forests

1.1 Measuring Heterogeneity without Machine Learning

To this point, our analysis has examined heterogeneity along various manually-defined subgroup dimensions. However, there is a small but rapidly growing literature in so-called "causal machine learning" which has provided promising methods to nonparamatrically characterize treatment effect heterogeneity in a data-driven manner (e.g. Athey and Imbens (2016), Wager and Athey (2018), Athey et al. (2019), etc.). While we have presented results from subgroups which one may theoretically expect to drive variation in treatment effect magnitudes across states and districts, there are many reasons why it may be beneficial to move beyond manual subgroup definitions into data-driven subgroup definitions.

Consider a simple, hypothetical scenario in which a researcher wishes to estimate the average treatment effect (ATE)[1] of a binary treatment, $W_i \in \{0, 1\}$, on an outcome, Y_i, where identification of this causal effect comes from the conditional independence assumption (CIA) – i.e. individual assortment into treatment status is random conditional (or unconditional) on a set of covariates, $\boldsymbol{X}_i$. Under the CIA, the researcher might estimate the following regression and obtain an unbiased estimate of the ATE, defined as the difference in potential outcomes for observation i under treatment vs. no treatment, $\tau = E[Y_i(1) - Y_i(0)|\boldsymbol{X}_i]$:

$$Y_i = \alpha + \tau W_i + \boldsymbol{X}_i'\boldsymbol{\beta} + \varepsilon_i \tag{1}$$

Now, the researcher wishes to understand how τ varies along the dimensions of $\boldsymbol{X}_i$. The usual approach to handling this question is to add an interaction term in the regression (assuming first, for simplicity, there is a single variable, X_i, in $\boldsymbol{X}_i$):

$$Y_i = \alpha + \tau W_i + \beta_1 X_i + \beta_2 W_i X_i + \varepsilon_i \tag{2}$$

After estimating (2), the researcher concludes that the treatment effect along the X_i dimension is $\tau + \beta_2 X_i$. While this approach may be satisfactory for a single dimension of heterogeneity, as $\boldsymbol{X}_i$ grows in its' dimensions, the number of parameters which must be estimated under this framework grows rapidly as the researcher must interact each dimension of heterogeneity with (1) W_i and (2) all other covariates in all possible combinations. To illustrate, if $\boldsymbol{X}_i$ contains just five covariates, there are $\sum_{x=1}^{5} C_x^5 = \binom{5}{1}+\binom{5}{2}+\binom{5}{3}+\binom{5}{4}+\binom{5}{5} = 31$ treatment-covariate interactions to estimate and another $\sum_{x=1}^{5} C_x^5 = 31$ covariate-covariate interactions to estimate. Including the ATE parameter itself and the intercept, the researcher must estimate a grand total of 64 parameters in this linear model. Figure 1 illustrates how rapidly the number of parameters grows with the dimensionality of $\boldsymbol{X}_i$. Even still, this linear model fails to consider non-linearities such as higher-order polynomials of the covariates and/or any important interactions between those polynomials.

[1] Under the CIA *and* homogeneity of treatment effects, the ATT = ATE. However, if there is treatment effect heterogeneity, then Equation (1) will identify the ATE, which will not equal the ATT. It is important to be aware of which treatment effect we are after when we generalize Equation (1).

In practice, it is likely that most interactions have little impact on treatment effect magnitude. Yet, OLS cannot inherently identify and place greater weights on more important covariates, yielding a traditional approach to heterogeneity mapping that is both cumbersome to interpret and dangerous to estimate, as the researcher loses statistical power quickly and runs the risk of spurious results due to multiple hypothesis testing.

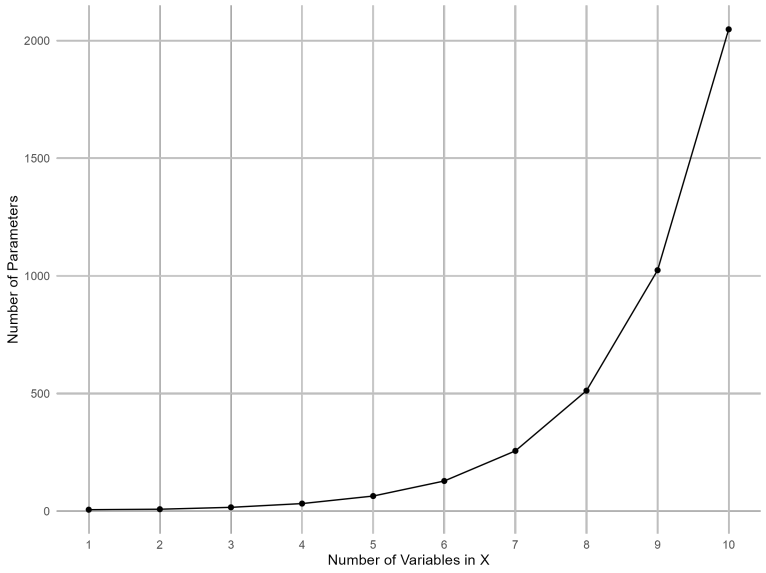


Figure 1: Parameters vs. Variables

Notes: This graph shows the number of parameters one must estimate in a fully-interacted linear model akin to Equation (2) as the number of covariates (i.e. possible dimensions of heterogeneity) in $\boldsymbol{X}_i$ increases.

Because of this "curse of dimensionality" problem and multiple hypothesis testing limitations, researchers often will either add one covariate interaction at-a-time (estimating a large number of regressions rather than a large number of parameters to retain statistical power in each model) or will discretize continuous variables along theoretical (or sometimes arbitrary) thresholds. While these methods certainly provide some information on the nature of heterogeneity in the data, there are two major limitations to these approaches. Firstly, they are not generally able to control for correlations among covariates. That is, colinearity between heterogeneity dimensions will yield similar results regarding their relative contribution to the average treatment effect, obscuring which covariate is driving the heterogeneity. For example, if school district income is heavily correlated with school spending, both may return significant coefficients in their respective regressions while it may be that school spending is the only true driver of the effect. The second limitation is that, even if there is good reason for the selection of thresholds when discretizing, user-sepcified partitions may be suboptimal. These limitations mean that (1) researchers will *only* find heterogeneity where they look and (2) researchers may misspecify the nature of the heterogeneity in the place they are looking.

1.2 Overview of Generalized Random Forests

In the prior subsection, we discussed the limitations of OLS when conducting heterogeneity analysis. The core of the issue is that OLS is well equipped to estimate average impacts, but poorly equipped to identify subgroups where average impacts are similar within-group. In this section, we seek to address this problem through a nonparametric mapping of the conditional average treatment effect function with generalized random forests (Wager and Athey (2018); Athey et al. (2019)). Generalized random forests (GRFs) are inherently matching algorithms. However, they do not employ matching for identification of the ATE but rather, they employ matching to identify *heterogeneity* in the ATE. This subsection will first provide a primer on the "causal tree" (Athey and Imbens, 2016), an adapted variant of the standard decision tree which has been shown to retain the beneficial asymptotic properties of OLS while estimating ATEs within

subgroups. Second, we outline the training procedure and the aggregation of causal trees into a "causal forest." Third, we discuss adapting the standard causal tree/forest algorithm to a difference-in-differences setting with continuous treatment dosage. Finally, we discuss the relative benefits of the GRF approach in terms of selection bias control within subgroups.

1.2.1 Causal Trees

The high-level idea of the causal tree is that one can leverage matching estimators to identify subgroups with common treatment effects when identification of the ATE itself comes from another source (e.g. randomized treatment assignment, instrumental variables, etc.). In order to illustrate how the causal tree acts as a matching estimator, let us consider the estimate of the ATE from Equation (1) expressed as a matching estimator with corresponding potential outcomes:

$$\begin{aligned}\hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n}[Y(W_i = 1|\boldsymbol{X}_i) - Y(W_i = 0|\boldsymbol{X}_i)] \\ &= E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 0] \\ &= \underbrace{E[Y_i(1) - Y_i(0)|W_i = 1]}_{\text{ATE}} + \underbrace{E[Y_i(0)|W_i = 1] - E[Y_i(0)|W_i = 0]}_{\text{Selection/Omitted Variable Bias}}\end{aligned}$$

If the conditional independence assumption (CIA) holds, then $\boldsymbol{X}_i$ represents the true set of confounders. This implies that $E[Y_i(0)|W_i = 1, \boldsymbol{X}_i] = E[Y_i(0)|W_i = 0, \boldsymbol{X}_i]$ and therefore, the estimate of $\hat{\tau}$ is unbiased. Intuitively, constructing a matching estimator with $\boldsymbol{X}_i$ to recover an unbiased estimate of τ works because we are comparing the outcomes of *similar enough* treatment and control units such that treatment assignment is random among that subset[2]. However, one need not employ matching for the purpose of constructing valid control groups if valid control groups already exist. Instead, the goal of the causal tree is to construct alternate *subgroups* where the covariates in $\boldsymbol{X}_i$ are used to match on *treatment effect* rather than treatment status probability. Therefore, the implicit goal of the causal tree and causal forest is to relax Equation (1) such that the ATE is a *function* for the *conditional* average treatment effect (CATE) and we seek to map that function:

$$Y_i = \alpha + \tau(\boldsymbol{X}_i)W_i + f(\boldsymbol{X}_i) + \varepsilon_i \tag{3}$$

$$\tau(\boldsymbol{X}_i) = E[Y_i(1) - Y_i(0)|\boldsymbol{X}_i] \tag{4}$$

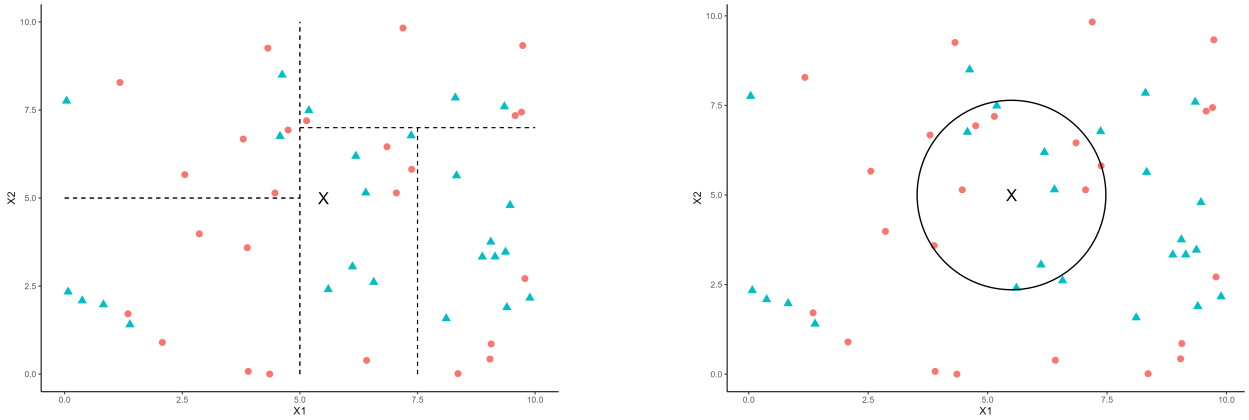
Causal trees group observations by greedily and recursively[3] splitting the data along covariates, $x_k \in \boldsymbol{X}$, and thresholds, c, into finer and finer subsamples. Figure 2a provides a simple example of this process, contrasted against an alternative matching algorithm based on researcher-specified Euclidean distance (2b), a common technique popularized in the corporate finance literature by Barber and Lyon (1996). Splits occur at "nodes." The resulting partitions from any given split are known as "children nodes" whereas the larger sample which was split is aptly called a "parent node." Further splits are made along children nodes into additional children nodes until stopping criteria are met, in which case we have a "leaf" (or sometimes called a "terminal node"). In total, there are 5 leaves in this sample tree. A sample test point $\mathbf{x}$ is shown, where its' treatment effect estimate will be the ATE of training points within its' leaf.

The optimal split at any given node is determined through gradient descent. The first step is to estimate the ATE within the parent node, $\hat{\tau}_P$, which is the GMM solution to a local moment condition implied by the identification strategy.[4] In the case of Equation (3), the root node estimate of $\hat{\tau}$ corresponds to the OLS

[2] This is true in both exact and inexact matching as it is well-known that controlling for "true" propensity scores, $p(\boldsymbol{X}_i)$ will yield a similar condition in that $E[Y_i(0)|W_i = 1, p(\boldsymbol{X}_i)] = E[Y_i(0)|W_i = 0, p(\boldsymbol{X}_i)]$

[3] "Greedy" splitting refers to the process of selecting the optimal split relative to the current node only: never selecting suboptimal splits or reverting back to prior splits. "Recursive" splitting refers to continuous splitting along the parent node, then subsequently along child nodes, and so on until stopping criteria are met.

[4] Extending the causal forest procedure to apply to any local moment condition is the work of Athey et al. (2019) in their generalized random forest procedure. We will use the term "causal forest" and "generalized random forest" interchangeably, but in all cases we are applying the generalized random forest as the causal forest is a special case of the GRF where treatment assignment is purely random.



(a) Example Tree-Based Matching

(b) Example Euclidean Matching

Figure 2: Tree-Based Matching vs. Euclidean Matching

Notes: The left figure shows an example of recursive partitioning (i.e. tree-based matching) in a simple case of two covariates. In this example, "treated" units are the blue triangles and "untreated" units are the red circles. The first split (i.e. "root node") partitioned the data along X_1, segmenting those units where $X_1 \geq 5$ in one "child node," C_1, and those with $X_1 < 5$ in the other (C_2). Within C_1, the algorithm again partitioned into two further children nodes (where C_1 is now itself a "parent" node) according to the threshold $X_2 = 7.5$. Similarly, in C_2, a further split was made according to $X_2 = 5$. These splits in each case were made to maximize the squared-difference in treatment effect estimates across the resulting child nodes. In total, this sample tree has 5 leaves (a.k.a. terminal nodes) where treatment effect predictions for test points within each leaf will be constant. The right figure shows a contrasting example of Euclidean-distance matching similar to Barber and Lyon (1996).

estimate of the ATE because the relevant local moment condition is the CIA and no splits have been made yet:

$$\begin{aligned} & E[W_i' \varepsilon_i | \boldsymbol{X}_i] = 0 \\ \implies & E[W_i'(Y_i - W_i \tau(\boldsymbol{X}_i)) | \boldsymbol{X}_i] = 0 \\ \implies & \frac{1}{n} \sum_i W_i Y_i - \frac{1}{n} \sum_i W_i^2 \hat{\tau}_P = 0 \\ \implies & \hat{\tau}_P = \frac{\sum_i W_i Y_i}{\sum_i W_i^2} \end{aligned} \tag{5}$$

The second step is to search over covariates and thresholds for the split that maximizes the squared-difference in estimated average treatment effects between the resulting partitions, C_1 and C_2, which can be described in the following objective function:

$$\max_{x_k, c} \quad n_{C_1} n_{C_2} (\hat{\tau}_{C_1} - \hat{\tau}_{C_2})^2 \tag{6}$$

Where $\hat{\tau}_{C_j}$ is approximated via a first-order Taylor series in order to remain computationally tractable:

$$\hat{\tau}_{C_j} \approx \hat{\tau}_P - \frac{\sum_{i \in C_j} W_i (Y_i - W_i \hat{\tau}_P)}{\sum_{i \in C_j} W_i^2} \tag{7}$$

Note that this can also be described in terms of the influence function, $\rho_{i,P}$, for observation i on the ATE in a given parent node, P:

$$\rho_{i,P} = \frac{W_i (Y_i - W_i \hat{\tau}_P)}{\frac{1}{n_P} \sum_{i \in P} W_i} \tag{8}$$

Where the objective function, Equation (6), is equivalent to maximizing the sum of the average influence functions between C_1 and C_2 as each average influence function capturess how much the treatment effect in the child node is likely to differ from the treatment effect estimate in the parent node. So Equation (6) may be restated as the following:

$$\max_{x_k, c} \quad \tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{n_{C_j}} \Big(\sum_{i \in C_j} \rho_{i,P} \Big)^2 \tag{9}$$

A key implementation detail is a process Athey and Imbens (2016) call "honesty." That is, two disjoint subsamples are used in the training of a single causal tree: one for determining the tree structure in the manner described above, and one for populating the nodes of the tree with treatment effect predictions. This is a critical step for ensuring treatment effect point estimates for out-of-sample test points are asymptotically normal and consistent, and eliminates the need for pruning.

Another point on implementation is regarding hard-coded splitting rules which help control balance, commonly referred to as "hyperparameters" in the machine learning literature. A potential problem which arises from "greedy" splitting is that unbalanced splits may occur (i.e. too few treatment or control units in a terminal leaf). These hyperparameters essentially define stopping rules such that a tree knows when to cease further splitting along a branch. Some of the more important ones include (1) the minimum number of treatment/control observations in a child node (2) the minimum share of the parent node that must be contained in a child node (3) the share of the training sample used as hold-out for honest estimation (4) an optional imbalance penaly parameter. The default hyperparameter settings in the 'grf' package in R (a stable CRAN package written by the original authors of the algorithm) are generally well-optimized for performance, though sometimes the original authors recommend more strict regularization when, for instance, there is more noise and/or less training data available.[5]

1.2.2 Aggregating Causal Trees into a Causal Forest

Wager and Athey (2018) note that, while point estimates of singluar causal trees have beneficial asymptotic properties, their structure remains quite sensitive to the initial split used to create the training data and they do not produce "smooth" estimates of the conditional average treatment effect (CATE) function, $\hat{\tau}(\boldsymbol{x})$. In order to improve out-of-sample prediction quality and generate a smoother function mapping, the causal forest algorithm grows B causal trees, where each tree is grown on a random sub-sample of the training data[6] and a random sub-sample of the covariate vector, $\boldsymbol{X}$. Aggregating across all trees, we can describe the CATE function as a weighted-average of the ATE, where the weight on training point i, α_i, at a given test point, $\boldsymbol{x}$, is the share of total trees where training point i falls into the same leaf, L_b, as $\boldsymbol{x}$:

$$\alpha_i(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{1}\{\boldsymbol{X}_i \in L_b(\boldsymbol{x})\}}{|L_b|} \tag{10}$$

Which yields the CATE function as estimated by the causal forest:

$$\hat{\tau}(\boldsymbol{x}) = \frac{\sum_i^n \alpha_i(\boldsymbol{x}) W_i Y_i}{\sum_i^n \alpha_i(\boldsymbol{x}) W_i^2} \tag{11}$$

We illustrate this process with two figures. First, Figure 3 illustrates how prediction quality improves and the functional mapping becomes smoother as more trees are grown and predictions are aggregated across trees. Here, data are generated according to a simple sine wave with additional idiosyncratic noise. We train 1000 decision trees in a random forest, plotting the first single tree as the light blue line and the forest-aggregated predictions as the dark blue line. The singular decision tree creates a piecewise functional mapping of the data generating process which is less accurate than the smoother mapping produced by the random forest.

The second illustrative figure is Figure 4, which displays how the causal forest aggregation creates a non-parametric kernel of training sample points relative to a test point with simulated data. In this example, there are three trees in a hypothetical forest, each creating their own partitions of the data with treated units represented as blue triangles and contol units as red circles, displayed in panels (a) – (c). Panel (d) shows

[5] See this GitHub post by one of the authors as an example: https://github.com/grf-labs/grf/issues/120#issuecomment-327276697

[6] Readers more familiar with random forests will note that this is slightly different from the bootstrapped sub-samples of the training data used in traditional random forests.

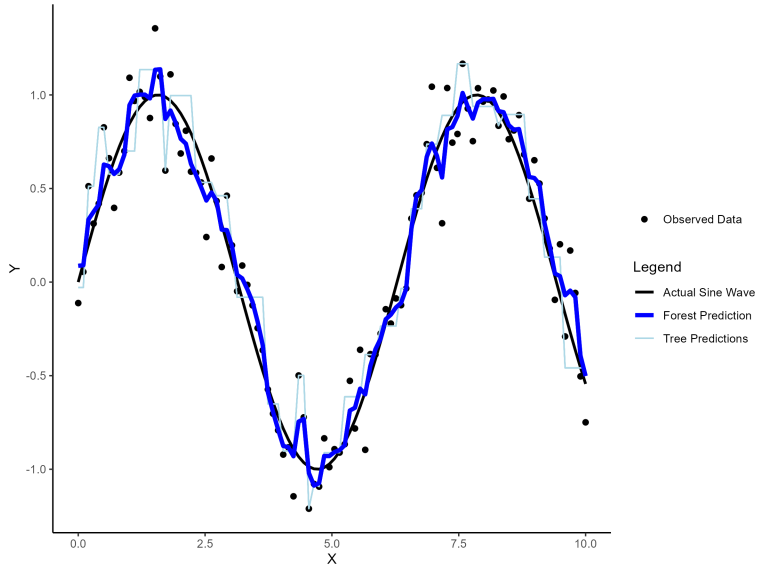


Figure 3: Smoothing and Prediction Improvement with Tree Aggregation

Notes: This graph shows 1 of 1000 regression trees designed to predict a sine wave data generating process and the resulting smoother mapping produced from aggregating predictions across all trees in a random forest.

the resulting weights on training points related to test point $\boldsymbol{x}$. Points which were in the same terminal leaf as $\boldsymbol{x}$ more frequently are given more weight in the prediction of $\boldsymbol{x}$'s individualized treatment effect.

1.2.3 A Brief Note on Centering

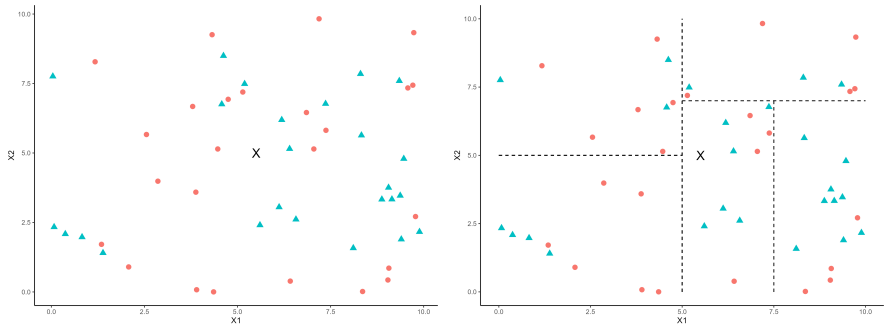
Before moving on, there is one additional implementation detail of the causal forest estimator we have omitted for simplicity. In practice, the GRF actually estimates the CATE using residualized (a.k.a. re-centered) observations to improve prediction quality. That is, the GRF estimates two auxilliary regression trees: one to predict the conditional mean function, $m(\boldsymbol{x}) = E[Y_i|\boldsymbol{X}_i]$, and another to predict the propensity score function, $e(\boldsymbol{x}) = E[W_i|\boldsymbol{X}_i]$. These auxilliary trees are trained via "leave-one-out" cross validation, denoted with the notation $(-i)$. That is, the tree uses all data except observation i in order to predict the conditional mean/score for observation i, leaving us with n trees trained on $n-1$ observations. The causal forest algorithm uses these estimates to recenter observations before proceeding with the training of causal trees and the subsequent forest aggregation. This produces a more robust and better performing estimate of the CATE function (Robinson (1988); Athey et al. (2019)):

$$\hat{\tau}(\boldsymbol{x}) = \frac{\sum_i^n \alpha_i(\boldsymbol{x})\widetilde{W}_i\widetilde{Y}_i}{\sum_i^n \alpha_i(\boldsymbol{x})\widetilde{W}_i^2} \tag{12}$$

$$\text{Where:} \quad \widetilde{Y}_i = Y_i - \hat{m}^{(-i)}(\boldsymbol{x}); \quad \widetilde{W}_i = W_i - \hat{e}^{(-i)}(\boldsymbol{x})$$

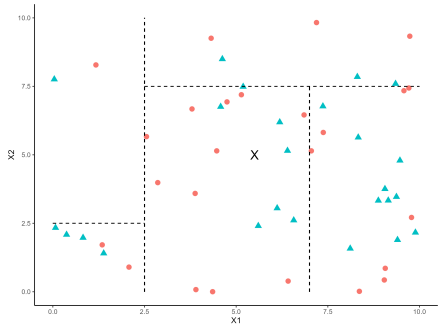
1.3 Adapting the GRF to a Difference-in-Differences Setting

The prior subsection provided a primer on the causal forest procedure for identifying the conditional average treatment effect when treatment assignment is random conditional (or unconditional) on a set of covariates. We noted that Athey et al. (2019) generalized this procedure such that the causal forest algorithm (in the form of generalized random forests – GRF) could be flexible enough to leverage any identification assumption which includes the ATE as a parameter, provided there is an accompanying local moment condition which can be solved via GMM. In this subsection, we outline our approach to applying this generalized algorithm to identify heterogeneous treatment effects in our panel data setting with continuous (dosage) treatment assignment.

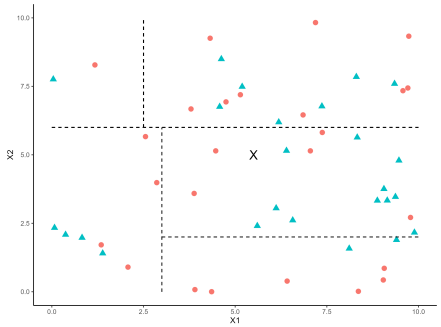


(a) Raw Data

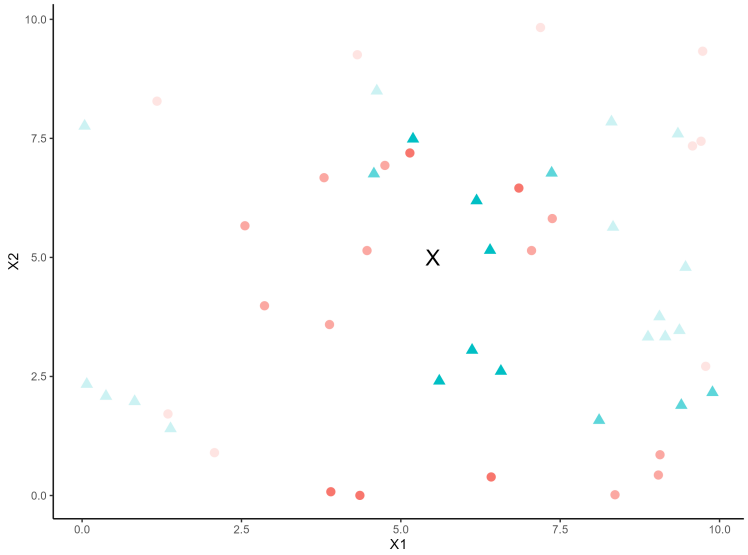
(b) Tree 1



(c) Tree 2



(d) Tree 3



(e) Forest-Defined Kernel

Figure 4: Illustration of Forest-Based Aggregation

Notes: These figures show an example derivation of forest-defined weights for a given test point, $\boldsymbol{x}$, in a hypothetical causal forest with 3 trees. Panels (b), (c), and (d) illustrate partitions created by each of the three trees, and panel (e) shows the aggregation and how points are re-weighted based on how frequently they fall in the same terminal leaf as $\boldsymbol{x}$.

1.3.1 Replacing the Conditional Independence Assumption with Parallel Trends

As discussed previously, the GRF procedure leverages the solutions of local moment conditions implied by the identification assumptions of the ATE. In a panel setting such as ours, the ATE is identified via the difference-in-differences estimator (DID) which notably relies on the parallel trends assumption (PTA) rather than the CIA. In this subsection, we will briefly review the PTA and, more importantly, discuss which specific version of the PTA from the recent DID literature we are applying in our GRF use-case.

Let us begin with the simple 2-group, 2-period case with binary treatment assignment. The basic PTA states that the observed change in outcomes for the control group between the pre- and post- treatment period (right-hand side) should equal the (counterfactual) change in outcomes for the treatment group between the pre- and post-treatment period had they not received treatment (left-hand side):

$$\text{PTA:} \quad E[Y_{i1}(0) - Y_{i0}(0)|W_i = 1] = E[Y_{i1}(0) - Y_{i0}(0)|W_i = 0]$$

If the PTA holds, we may derive the DID estimator of the ATE using observable moments:

$$\begin{aligned}\tau &= E[Y_{i1}(1)|W_i = 1] - E[Y_{i1}(0)|W_i = 1] \\ &= E[Y_{i1}(1)|W_i = 1] - \Big(E[Y_{i1}(0)|W_i = 0] - E[Y_{i0}(0)|W_i = 1] + E[Y_{i0}(0)|W_i = 1]\Big) \\ &= E[Y_{i1}(1) - Y_{i0}(0)|W_i = 1] - E[Y_{i1}(0) - Y_{i0}(0)|W_i = 0]\end{aligned} \tag{13}$$

The most common method of estimating τ from Equation (13) is via a regression specification which includes unit and time fixed effects, along with a treatment indicator for if unit $i = 1, ..., N$ is treated in period $t = 1, ..., T$:

$$Y_{it} = \gamma_i + \delta_t + \tau W_{it} + \varepsilon_{it} \tag{14}$$

Equation (14) is a general specification which allows consistent and asymptotically normal estimation of τ under the PTA, no-anticipation assumption, and independently sampled units (Roth et al., 2023). Therefore, Equation (14) implicitly assumes that parallel trends hold conditional on unit and time fixed effects. One of the advantages of this approach is that it allows for the identification of an average causal effect in the presence of treatment effect heterogeneity stemming from time-invariant covariates and/or time variant covariates which affect all units equally because unit and time fixed effects absorb each of these impacts, respectively.

In the more common scenario, researchers should consider the "staggered" parallel trends assumption (we borrow the notation in this section from Roth (2022)). There are $t = 1, ..., T$ time periods and units can be treated in any period after the first, $t > 1$, and they must remain treated for the duration of the panel. Denote the first period where unit i receives treatment as G_i and let $D_{i,t}$ be an indicator for if unit i receives treatment in period t – so G_i is just an index for the earliest period where $D_{i,t} = 1$. Note that control units (never treated) have $G_i = \infty$.

We must refine the potential outcomes for unit i in this case to allow for flexibility in treatment timing. Define a vector of 0's and 1's corresponding to the potential outcomes of being untreated or treated in any given period as $\mathbf{0}_s$ and $\mathbf{1}_s$, which are of some flexible length s. If unit i is first treated in some period $G_i = g$, then their potential outcome in period t is what their outcome would be from not having received treatment in period $g - 1$ followed by receiving treatment in every period from g to t ($T - g - 1$): $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g-1})$. Conversely, their potential outcome from being never-treated just uses all 0's for the panel duration: $Y_{i,t}(\mathbf{0}_T)$. Since units that are treated remain treated, we can simplify the notation a bit:

- Denote the potential outcomes for all units first treated in period g as: $Y_{it}(g)$
- Denote never-treated units as: $Y_{it}(\infty)$

The Staggered PTA (SPTA) states that the average outcomes for all treated groups would have evolved in parallel between any two time periods t and t' in the absence of treatment:

$$\begin{aligned}\text{SPTA:} \quad E[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g] = E[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g'], \\ \forall t \neq t',\ g \neq g'\end{aligned} \tag{15}$$

Where the LHS above represents the change in outcomes for the group treated in period g between any two periods t and t' had they never been treated, and the RHS represents the change in outcomes for all other groups (i.e. those treated in any other period *and* those never treated) between any two time periods had they not been treated.

This is a restrictive assumption which is unlikely to hold in practice. So Sun and Abraham (2021) and Callaway and Sant'Anna (2021) relax the SPTA a bit in two ways. They are able to derive estimators which require only that:

1. Trends must be parallel between some period t and the last period before treatment $g-1$ (denoted g_{min}).
2. Trends must be parallel between groups that are eventually treated – i.e. trends don't need to be parallel between eventually treated and never-treated units.

Their adjusted SPTA is the following:

$$\begin{aligned} \text{SPTA}^{adj}: \quad & E[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g] = E[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g'], \\ & \forall t \neq t',\ g \neq g'; t, t' \geq g_{min} \end{aligned} \tag{16}$$

Note that SPTA^{adj} is largely the same, only it specifies that this is only for periods after g_{min} and now g' no longer includes never-treated units where $g' = \infty$.

Now we may estimate τ under SPTA^{adj} by modifying Equation (14) so that we interact the indicator for $(Treat \times Post)$, W_{it}, with another indicator for *relative* treatment time: $\mathbb{1}[R_{it}]$. Here, W_{it} acts as normal: it equals 1 if the unit is treated in period t. And now we add $\mathbb{1}[R_{it}] = T - g_i + 1 = r$ so that it equals 1 in the first period after treatment, 2 in the next, etc:

$$Y_{it} = \gamma_i + \delta_t + \sum_{r \neq 0}^{T*} \tau W_{it} \cdot \mathbb{1}[R_{it} = r] + \varepsilon_{it} \tag{17}$$

What $\mathbb{1}[R_{it}]$ means is that if, say, $R_{it} = 2$ (i.e. we are in the second treatment period for unit i), then the indicator equals 1. So we are estimating τ separately for every relative time period after treatment (i.e. we are running many different regressions, 1 for each relative time period). Under the SPTA^{adj} (and no anticipation/independent sampling), this version of TWFE will produce consistent estimates of τ. However, Sun and Abraham (2021) show that if there is heterogeneity in effects which varies by adoption timing (a problem brought up earlier), the estimates produced by the above equation will be hard to interpret because of "cross-contamination" and "negative weighting."

Callaway and Sant'Anna (2021) propose an estimator which tries to overcome this issue. They look at (1) the change in outcomes for treated units from a "base year" – usually period $t_{base} = g_i - 1$, the period right before initial treatment – to year t, and compare to (2) the change in outcomes over the same period for never-treated and not-yet-treated individuals. They estimate treatment effects separately for every combination of (1) treatment timing and (2) relative time after treatment.

Note that there is another type of estimator which tries to address this issue implemented by Borusyak et al. (2024). Their approach is to fit a TWFE regression on all not-yet-treated units, and impute the predicted outcomes of not-yet-treated units in period t as the counterfactual for treated units in period t. This approach yields consistent estimates of τ under the stronger version of the SPTA (i.e. PTA must hold for all time periods and units). This approach has the advantage of being more efficient than Callaway and Sant'Anna (2021) – especially if there is anticipation, but is more prone to bias in the event of diverging trends (e.g. if group-specific linear trends are present) or if there is serial correlation in the errors.

This brings us to our use-case: how the DID identification assumptions change when the treatment effect is a function of unit-specific, time-invariant characteristics. Let us now allow the ATE as expressed in

Equation (14) to vary individually according to a set of time-invariant covariates, $\boldsymbol{X}_i$:

$$Y_{it} = \gamma_i + \delta_t + \tau(\boldsymbol{X}_i)W_{it} + \varepsilon_{it} \tag{18}$$

In the case of Equation (18), the impacts of $\boldsymbol{X}_i$ on Y_{it} cannot be identified because they are obsorbed in γ_i. However, Kattenberg et al. (2023) discuss that in principle, if one could identify subgroups, S, of individuals where the treatment effect is nearly identical, $\tau(\boldsymbol{X}_i) \approx \tau \quad \forall i \in S$, then one could use the estimators of Borusyak et al. (2024) or Sun and Abraham (2021) to isolate this effect for members of that subgroup under a *localized* staggered parallel trends assumption. This local SPTA is effectively just the Sun and Abraham (2021) SPTA^{adj}, with the only additional requirement being that it should hold for units within subgroup S:

$$\begin{aligned} \text{SPTA}^{local}: \quad & E[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g] = E[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g'], \\ & \forall t \neq t',\ g \neq g';\ t, t' \geq g_{min};\ \forall i \in S \end{aligned} \tag{19}$$

1.3.2 Causal Forest with Fixed Effects

Kattenberg et al. (2023) outline a computationally tractable approach for using generalized random forests to define subgroups in a DID setting. To build intuition, consider a simple case where there are only two subgroups with unique treatment effects, $g = \{0, 1\}$, of size N^g. Denote their group-specific treatment effect as τ^g. Of course, if we knew the subgroups ex-ante, then we could simply estimate τ^g by restricting the sample to only those units in group g and estimating Equation (18) or by including an interaction term in Equation (18) representing a group identifier along with the treatment indicator:

$$\begin{aligned} Y_{it} = \gamma_i + \delta_t + \tau^g W_{it} + \varepsilon_{it}; \\ \forall i \in g; \quad \forall g \in \{0, 1\} \end{aligned} \tag{20}$$

This is exactly the approach we conduct in the first sections of this paper with theoretical subgroups. At this point, the GRF procedure is supposed to enter: we apply the non-parametric search procedure for subgroups with similar treatment effects and derive a weighting kernel that allows a mapping of the conditional average treatment effect function, $\tau(\boldsymbol{x})$. However, recall that the GRF estimates auxilliary regression trees for the propensity score function, $e(\boldsymbol{x}) = E[W_{it}|\boldsymbol{x}, \gamma_i, \delta_i]$, and the conditional mean function, $m(\boldsymbol{x}) = E[Y_{it}|\boldsymbol{x}, \gamma_i, \delta_i]$. This poses a problem because the propensity scores will be clustered around 1 or 0 with the inclusion of time and unit fixed effects, which violates the identifying assumptions of the GRF. Therefore, it is necessary to modify Equation (20) with a double within-transformation (see Wooldridge (2010) for a detailed discussion) such that τ^g is estimated in a numerically equivalent manner without the use of fixed effects. Specifically, we first average within-unit (i.e. over time) and obtain the cross section equation:

$$\bar{Y}_i = \tau^g \bar{W}_i + \gamma_i + \bar{\varepsilon}_i \tag{20a}$$

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}; \quad \bar{W}_i = \frac{1}{T}\sum_{t=1}^{T} W_{it}$$

Subtracting (20a) from (20) yields the first transformation and eliminates unit-specific fixed effects:

$$\ddot{Y}_{it} = \delta_t + \tau^g \ddot{W}_i + \ddot{\varepsilon}_i \tag{20b}$$

$$\ddot{Y}_{it} = Y_{it} - \bar{Y}_{it}; \quad \ddot{W}_{it} = W_{it} - \bar{W}_{it}; \quad \ddot{\varepsilon}_{it} = \varepsilon_{it} - \bar{\varepsilon}_{it}$$

Next, we average within-year (i.e. over units) and create another transformed version of Equation (20) which eliminates time fixed effects. Notice the critical detail in that we must use only observations in group g for this transformation. That is, it must be done locally:

$$\widetilde{Y}_{it} = \tau^g \widetilde{W}_{it} + \varepsilon_{it} \tag{20c}$$

$$\widetilde{Y}_{it} = \ddot{Y}_{it} - \frac{1}{N^g}\sum_{i \in N^g} \ddot{Y}_{it}; \quad \widetilde{W}_{it} = \ddot{W}_{it} - \frac{1}{N^g}\sum_{i \in N^g} \ddot{W}_{it}$$

Estimating (20c) with OLS yields a per-group estimate of the ATE which is equivalent to that obtained by estimating Equation (20):

$$\hat{\tau}^g = \frac{\sum_t \sum_{i \in g} \widetilde{W}_{it} \widetilde{Y}_{it}}{\sum_t \sum_{i \in g} \widetilde{W}_{it}^2}; \quad g = 0, 1 \tag{22}$$

Kattenberg et al. (2023) discuss that performing this double transformation manually (i.e. once before any splitting is done) and subsequently training a GRF on the transformed variables will produce a biased estimate of the conditional average treatment effect. This is because the second transformation must be conducted locally among units with similar treatment effects in order for the asymptotic properties of the GRF to continue to hold after recentering. Therefore, they propose recentering at individual nodes before searching for a split. Their procedure may be summarized in Algorithm 1:

Algorithm 1 Causal Forest Fixed Effects (CFFE)

1: **for** Each causal tree $b \in B$ **do**

2: Perform global recentering at root node, P. Seek the heterogeneity-maximizing split among covariates $x_k \in \boldsymbol{X}$ and thresholds c:

$$\widetilde{Y}_{it} = \ddot{Y}_{it} - \frac{1}{N} \sum_{i \in N} \ddot{Y}_{it}; \quad \widetilde{W}_{it} = \ddot{W}_{it} - \frac{1}{N} \sum_{i \in N} \ddot{W}_{it} \tag{1a}$$

$$\max_{x_k, c} \quad (\hat{\tau}_{C_1} - \hat{\tau}_{C_2})^2 \tag{1b}$$

3: At the first child nodes, C_1 and C_2, perform *local* recentering on the raw outcome variable and treatment indicator. Then seek the recursive heterogeneity-maximizing split:

$$\widetilde{Y}_{it} = \ddot{Y}_{it} - \frac{1}{N^{C_i}} \sum_{i \in N} \ddot{Y}_{it}; \quad \widetilde{W}_{it} = \ddot{W}_{it} - \frac{1}{N^{C_i}} \sum_{i \in N} \ddot{W}_{it} \quad \forall i \in \{1, 2\} \tag{2a}$$

$$\max_{x_k, c} \quad (\hat{\tau}_{C_1} - \hat{\tau}_{C_2})^2 \tag{2b}$$

4: Repeat (2a) and (2b) until stopping criteria are met.

5: **end for**

6: Produce the CATE function mapping by aggregating over trees:

$$\hat{\tau}(\boldsymbol{x}) = \frac{\sum_t \sum_i \alpha_i(\boldsymbol{x}) \tilde{W}_i \tilde{y}_i}{\sum_t \sum_i \alpha_i(\boldsymbol{x}) \tilde{W}_i^2} \tag{3}$$

1.4 Results

At this stage, we are not yet able to conduct estimation via the CFFE approach. Therefore, we present causal forest estimates from globally transformed data, with the major caveat that these estimates are primarily illustrative at this stage as they are expected to be biased absent of local recentering.

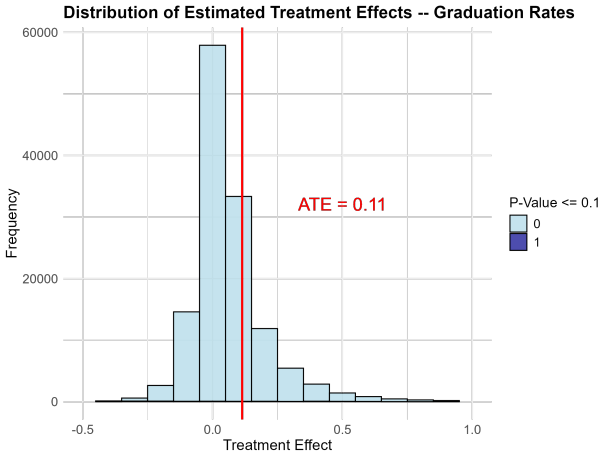
1.4.1 Distributions of Average Treatment Effects

The aforementioned estimation procedure produces individualized treatment effect (ITE) estimates at the school district × year level, which are equivalent to conditional average partial effects among the most granular subroup (the individual school district × year). Figure 5 displays this distribution along with the GRF-estimated average treatment effect.[7] This exercise produces ATE estimates which are qualitatively

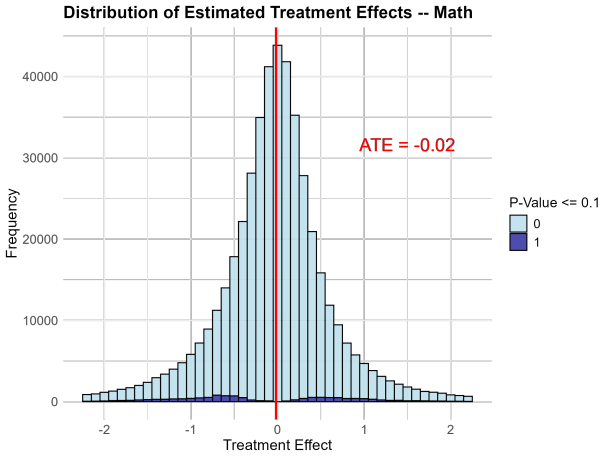
[7] Recall that the ATE will be the estimate of the subgroup ATE at the root node (that is, before any splits are made). So this estimate simply comes from OLS on the double-differenced $\tilde{y}_i$ and $\tilde{W}_i$. Since this numerically equivalent to the generalized DID estimator, this value should be very similar to the ATE estimated from that regression.

similar to those estimated by the GDD. For graduation rates, the causal forest estimates a slightly larger ATE of 0.11 compared to the GDD estimate of 0.076 – 0.089 in Sample 1; for math scores, the causal forest estimates an ATE of -0.02, nearly equivalent to that of the GDD estimate which ranged between -0.024 – -0.029; and for ELA scores, the causal forest estimated an ATE of 0.09, slightly larger than the GDD which estimated between 0.056 – 0.058. .

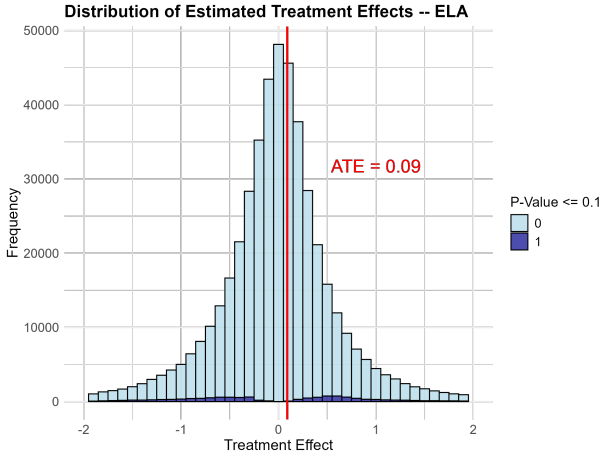
Figure 5 also presents evidence for heterogeneous treatment effects for all outcomes. While the majority of district × year treatment effect estimates are imprecise, there are a fair number of ITEs statistically different from zero, which we will explore in greater detail in the following subsections.



(a) Graduation Rates



(b) Math Scores



(c) ELA Scores

Figure 5: Distributions of District × Year ITE Estimates

Notes: These figures are the distributions of individualized (district × year)-level treatment effects for each of the three main outcomes from the causal forests. We estimate standard errors for each ITE, and shade darker those districts with statistically significant (at the $p = 0.1$ level) ITE estimates. We also display the GRF-estimated ATE which comes from the aggregation of doubly-robust scores. These are interpreted as average partial effects for a given district × year. That is, each point represents $E\left[\frac{\partial \tau(\boldsymbol{X}_{it})}{\partial W_{it}}\right] = \frac{Cov[Y_{it}, W_{it}|\boldsymbol{X}_{it}=\boldsymbol{x}]}{Var[W_{it}|\boldsymbol{X}_{it}=\boldsymbol{x}]}$, the predicted treatment effect from increasing the charter share in district d in year t by 1 percentage point.

1.4.2 Treatment Effects within Theoretical Subgroups

In order to best understand the pattern of heterogeneity in charter school impacts, we first aggregate the ITE estimates from Figure 5 into informative well-defined subgoups. Recall that the GRF estimation procedure, at its' core, produces a mapping of the heterogeneous treatment effect function. Therefore, we

can investigate the empirical importance of a theoretical subgroup by aggregating ITEs accordingly. Table 1 displays the results of this exercise. Here, we study the group average treatment effect (GATE), and corresponding statistical significance, of urban vs. suburban vs. rural districts, districts with a greater share of free or reduced-price lunch students, states with no cap laws on the maximum charter share, states with transprent application processes for initial charter application and renwewal, states with performance-based renewal contracts, and states with equitable funding for TPS vs. charter schools. A few interesting findings include: suburban and rural districts are more likely to see improved graduation rates with expanded charter schooling; states with performance-based contract renewal tend to see greater returns on graduation rates, but negative returns to math scores; and states with equitable funding for charter schools tend to see greater returns to both graduation rates and math scores.

Table 1: GATEs within Subgroups

	GATE Estimates			Proportion of N		
Group	Grad Rate	Math	ELA	Grad Rate Share	Math Share	ELA Share
Urban	-0.112 (0.123)	0.014 (0.052)	-0.056 (0.049)	0.060	0.057	0.057
Suburban	0.058** (0.029)	-0.011 (0.077)	-0.068 (0.068)	0.226	0.269	0.269
Rural	0.077* (0.04)	-0.095 (0.122)	-0.179 (0.163)	0.529	0.462	0.462
Percent Free Lunch > 30%	0.037 (0.041)	-0.015 (0.063)	-0.048 (0.048)	0.400	0.771	0.771
No Caps on CS Growth	0.003 (0.033)	0.213 (0.149)	-0.279 (0.223)	0.280	0.290	0.290
Trans. CS Startup Policies	0.037 (0.029)	0.207** (0.103)	-0.161** (0.079)	0.366	0.296	0.296
Performance-Based Contracts	0.053** (0.025)	-0.147*** (0.055)	-0.038 (0.052)	0.424	0.310	0.310
Equitable Funding	0.063*** (0.02)	0.303** (0.138)	-0.258 (0.185)	0.417	0.388	0.388

1.4.3 Covariate Profiles of Positively-Impacted vs. Negatively-Impacted Districts

The next step in the investigation examines the covariate profile of districts with significantly positive impacts contrasted with the covariate profile of those districts with significantly negative impacts. Tables 2 – 4 display the differences across these district profiles. Columns (1) and (2) show the covariate averages for districts with positive or negative treatment effects, respectively, and column (3) shows the difference in those covariate averages.

In Table 2, we see that districts with significantly positive ITE estimates for graduation rates tend to have greater enrollment, more white and Hispanic students, fewer black students, fewer students on free/reduced price lunch, more students in special ed, etc. Table 3 displays group averages for positive vs. negative math score districts. Districts with significantly positive impacts tend to have more Hispanic students, spend less at baseline per-student, have higher student-teacher ratios, have lower teacher salaries at baseline, have fewer magnet schools, performed worse at baseline, etc. Table 4 paints a somewhat similar picture for ELA scores: positively impacted districts spent less per-student at baseline, performed worse at baseline, had lower teacher salaries, etc. Distilling these results seems to suggest that test scores are more likely to be improved in districts where there is more room *for* improvement. That is, lower income districts, districts with higher student-teacher ratios, districts without many magnet schools, and districts performing worse at baseline tend to benefit more from charter school expansion.

Table 2: Comparing Covariate Means: Graduation Rates

Covariate	Significantly Positive	Significantly Negative	Difference (Positive - Negative)
Log of Enrollment	7.23	6.98	0.26**
White (%)	0.82	0.48	0.35***
Black (%)	0.05	0.39	-0.34***
Hispanic (%)	0.09	0.05	0.04***
Free/Reduced Lunch (%)	0.25	0.48	-0.23***
Special Ed (%)	0.26	0.13	0.13***
Baseline Performance	0.83	0.57	0.26***
Urban	0.07	0.07	0
Suburb	0.22	0.14	0.08***
Town	0.22	0.17	0.06*
Rural	0.48	0.62	-0.14***
Magnet Schools (%)	0.18	0.01	0.17
City Population (standardized)	0.00	0.00	0
University in City	-0.01	0.01	-0.01**
Per Pupil Revenue	11411.08	8940.67	2470.41***
Student-Teacher Ratio	14.20	14.93	-0.73***
Teacher Salary	81975.53	65372.43	16603.09***
TPS-Charter Spending (% diff)	0.03	0.01	0.02***
Total Spending (per-pupil)	11480.57	9178.68	2301.89***
Equitable Funding	6.21	5.06	1.15***
No Caps on CS Growth	8.26	9.12	-0.86***
Performance-Based Contracts	8.31	9.75	-1.44***
Transparent Charter Startup Policies	7.98	10.16	-2.18***
Clear Charter Renewal Policies	10.54	11.91	-1.37***
Exempt from State/District Regs	7.51	8.04	-0.53***
Number of Observations	438.00	277.00	715

1.4.4 Variable Importance Factors and their Linear Relationship with Treatment Effects

The covariate profiles of (significantly) positively impacted vs. negatively impacted districts is an important first-step in the causal forest analysis of heterogeneity. With it, we have a profile for the "average" district that stands to benefit or receive harm from charter expansion. However, the covariate profile fails to answer exactly *which* covariates are driving the heterogeneity, and if the "direction" of the effect is linear, interactive, or other. For the next step in the analysis of heterogeneity, it is useful to look within the training procedure of the GRF for the covariates which proved to be the most informative when creating subgroups and, therefore, mapping the CATE function. Examining the share of trees which split along each covariate, weighted by the depth of the split itself (such that earlier splits are weighted more heavily) provides a rough approximation of which covariates influence the gradient of the CATE function. This measure is referred to in Wager and Athey (2018) as the "Variable Importance Factor," and we display these results in Figure 6.[8] We also explore these "important" covariates' linear relationship with the treatment effect estimate in Tables 5 – 7. This is what Athey et al. (2019) refer to as the "Best Linear Projection" of the CATE function. To create this, we simply regress the vector of covariates, $\boldsymbol{X}_{it}$, on the ITE prediction itself. This allows us to see the linear relationship between covariates and the treatment effect. We consider only the top 7 covariates according to VIF.

We find a similar set of covariates drive heterogeneity for all outcomes. Baseline performance tends to be a strong predictor for math and ELA scores, but is by far the strongest predictor of graduation rates. We

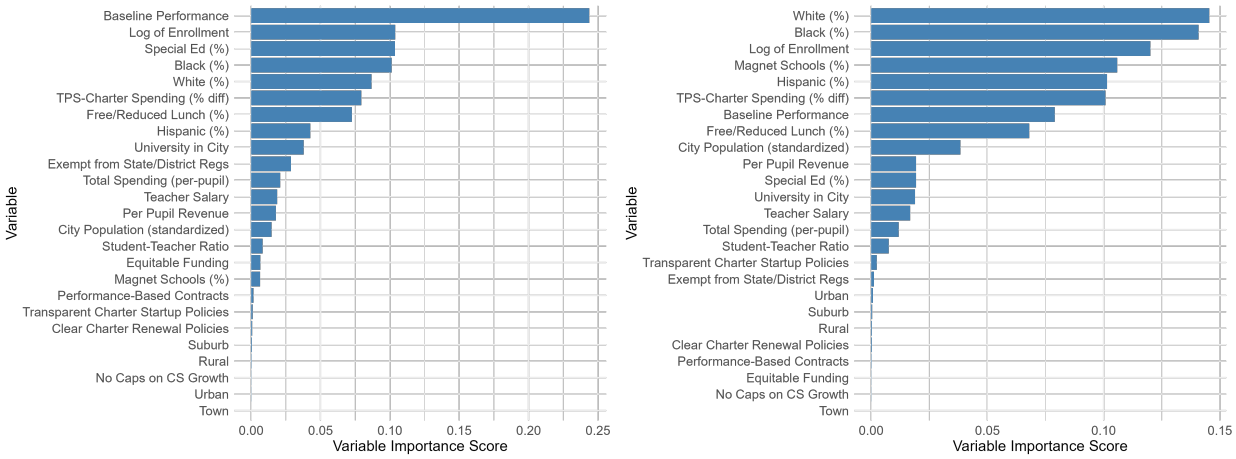
[8]The formula for variable importance for covariate x_k is: $VIF(x_k) = \sum_j \frac{1}{j^2}\left[\frac{\sum_{trees} \#\# \text{ depth-}j \text{ splits on } x_k}{\sum_{trees} \#\text{ total depth-}j \text{ splits}}\right]$

Table 3: Comparing Covariate Means: Math

Covariate	Significantly Positive	Significantly Negative	Difference (Positive - Negative)
Log of Enrollment	7.29	7.26	0.03
White (%)	0.74	0.78	-0.04***
Black (%)	0.07	0.06	0**
Hispanic (%)	0.12	0.09	0.03***
Free/Reduced Lunch (%)	0.45	0.42	0.04***
Special Ed (%)	0.14	0.15	-0.01***
Baseline Performance	0.28	0.40	-0.12***
Urban	0.03	0.04	0
Suburb	0.29	0.31	-0.02*
Town	0.17	0.16	0.02***
Rural	0.50	0.50	0
Magnet Schools (%)	0.09	0.34	-0.25***
City Population (standardized)	0.00	0.00	0
University in City	0.00	0.00	0***
Per Pupil Revenue	13857.58	14541.93	-684.35***
Student-Teacher Ratio	14.77	14.23	0.54***
Teacher Salary	95478.39	96551.46	-1073.08***
TPS-Charter Spending (% diff)	0.01	0.00	0.01***
Total Spending (per-pupil)	13697.81	14348.39	-650.59***
Equitable Funding	5.54	5.44	0.1*
No Caps on CS Growth	8.59	8.44	0.15***
Performance-Based Contracts	8.44	8.55	-0.11**
Transparent Charter Startup Policies	8.39	8.68	-0.29***
Clear Charter Renewal Policies	10.51	10.37	0.13***
Exempt from State/District Regs	7.21	6.93	0.28***
Number of Observations	6002.00	7786.00	13788

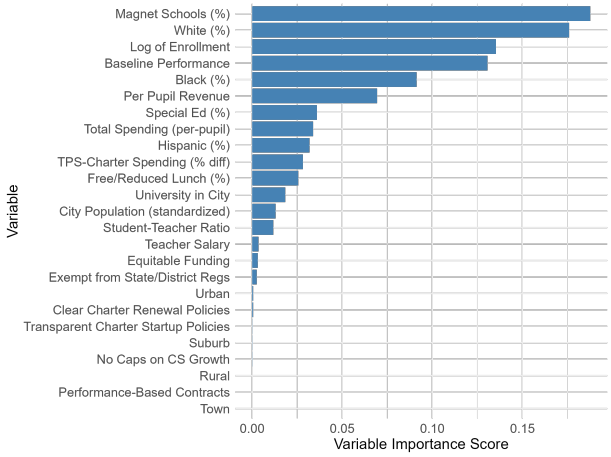
also find that the baseline enrollment level consistently appears in the top 3 predictors for each outcome. For test scores, the number of magnet schools in a district, as well as the share of white vs. black vs. Hispanic students tend to be important predictors. This information on relative importance combined with the covariate profiles in Tables 2 – 4 help us understand the nature of the heterogeneity. For graduation rates: higher levels of prior enrollment, higher levels of prior performance, a larger share of special ed students, fewer black students, and more white students are the main drivers of positive treatment effects. For math and ELA scores, fewer white students and more Hispanic students, fewer magnet schools, and lower performance at baseline are the drivers of positive impacts. Notice that the log of prior enrollment is always a strong predictor of treatment effects according to VIF, but does not differ meaningfully across positive vs. negative districts in either math or ELA scores. To understand why this may occur, recall that VIF is calculated as the depth-weighted share of the number of trees that split along a given covariate. Thus, if a covariate is used across many trees as an initial splitting variable, it is therefore necessarily informative as an interactor with other covariates. This is a non-linearity which is difficult to see from simply examining covariate means.

The BLP analysis investigates if there are any approximately linear relationships between important covariates and treatment effects. Above, we mentioned that important covariates may be important non-linearly or interactively with other covariates. In these cases, we cannot answer if "more of x is usually better/worse for $\tau(x)$," because the CATE function is of high dimensionality and non-convexity. But in some cases a somewhat linear relationship may exist, and the BLP allows us to explore that possibility. In Tables 5 – 7, we find little evidence for linear relationships. Only Baseline Performance and Baseline Enrollment have statistically significant positive linear relationships with the treatment effect predictions,



(a) Graduation Rates

(b) Math Scores



(c) ELA Scores

Figure 6: VIF Scores

Notes: These figures display variable importance factors (VIFs), which amount to the share of trees which split along a given covariate, weighted by the depth at which the split occured.

Table 4: Comparing Covariate Means: ELA

Covariate	Significantly Positive	Significantly Negative	Difference (Positive - Negative)
Log of Enrollment	7.30	7.32	-0.02
White (%)	0.77	0.74	0.03***
Black (%)	0.06	0.06	0
Hispanic (%)	0.10	0.12	-0.02***
Free/Reduced Lunch (%)	0.45	0.42	0.03***
Special Ed (%)	0.15	0.15	-0.01***
Baseline Performance	0.31	0.42	-0.11***
Urban	0.03	0.06	-0.03***
Suburb	0.27	0.35	-0.09***
Town	0.18	0.14	0.04***
Rural	0.52	0.45	0.07***
Magnet Schools (%)	0.07	0.29	-0.22***
City Population (standardized)	0.00	0.01	-0.01***
University in City	0.00	-0.01	0.01***
Per Pupil Revenue	13411.55	15221.81	-1810.26***
Student-Teacher Ratio	14.88	14.23	0.66***
Teacher Salary	92433.94	100392.11	-7958.17***
TPS-Charter Spending (% diff)	0.01	0.01	0
Total Spending (per-pupil)	13194.43	15038.02	-1843.59***
Equitable Funding	5.41	5.57	-0.16***
No Caps on CS Growth	8.55	8.54	0.02
Performance-Based Contracts	8.79	8.43	0.35***
Transparent Charter Startup Policies	8.39	8.60	-0.21***
Clear Charter Renewal Policies	10.60	10.45	0.15***
Exempt from State/District Regs	7.33	7.01	0.32***
Number of Observations	6454.00	7172.00	13626

and only for the graduation rate outcome.

Table 5: Best Linear Projection: Graduation Rates

Term	Estimate	Std. Error	t-stat	p-value
(Intercept)	-1.847**	0.721	-2.563	0.010
Baseline Performance	0.932**	0.380	2.453	0.014
Log of Enrollment	0.105**	0.047	2.247	0.025
Special Ed (%)	4.19	3.087	1.358	0.175
Black (%)	-0.898	0.667	-1.347	0.178
White (%)	-0.178	0.384	-0.463	0.643
TPS-Charter Spending (% diff)	-0.033	0.282	-0.116	0.908
Free/Reduced Lunch (%)	0.296	0.306	0.966	0.334
Hispanic (%)	0.157	0.423	0.371	0.711
University in City	-0.53*	0.317	-1.673	0.094
Exempt from State/District Regs	-0.019	0.034	-0.561	0.575

Table 6: Best Linear Projection: Math Scores

Term	Estimate	Std. Error	t-stat	p-value
(Intercept)	1.389	0.911	1.525	0.127
White (%)	-0.186	0.485	-0.384	0.701
Black (%)	0.097	0.568	0.170	0.865
Log of Enrollment	-0.115*	0.065	-1.775	0.076
Magnet Schools (%)	0.007	0.005	1.433	0.152
Hispanic (%)	0.082	0.450	0.182	0.856
TPS-Charter Spending (% diff)	0.211	0.193	1.093	0.274
Baseline Performance	0.079	0.084	0.939	0.348
Free/Reduced Lunch (%)	0.024	0.428	0.056	0.955
City Population (standardized)	0.177*	0.096	1.843	0.065
Per Pupil Revenue	0	0.000	-1.396	0.163

Table 7: Best Linear Projection: ELA Scores

Term	Estimate	Std. Error	t-stat	p-value
(Intercept)	0.969**	0.395	2.450	0.014
Magnet Schools (%)	0.004	0.007	0.630	0.528
White (%)	-0.58**	0.229	-2.532	0.011
Log of Enrollment	-0.008	0.032	-0.247	0.805
Baseline Performance	-0.017	0.047	-0.359	0.720
Black (%)	-0.576**	0.246	-2.342	0.019
Per Pupil Revenue	0	0.000	-1.112	0.266
Special Ed (%)	-2.143	1.328	-1.614	0.107
Total Spending (per-pupil)	0	0.000	0.627	0.530
Hispanic (%)	-0.745***	0.237	-3.148	0.002
TPS-Charter Spending (% diff)	0.301***	0.109	2.772	0.006

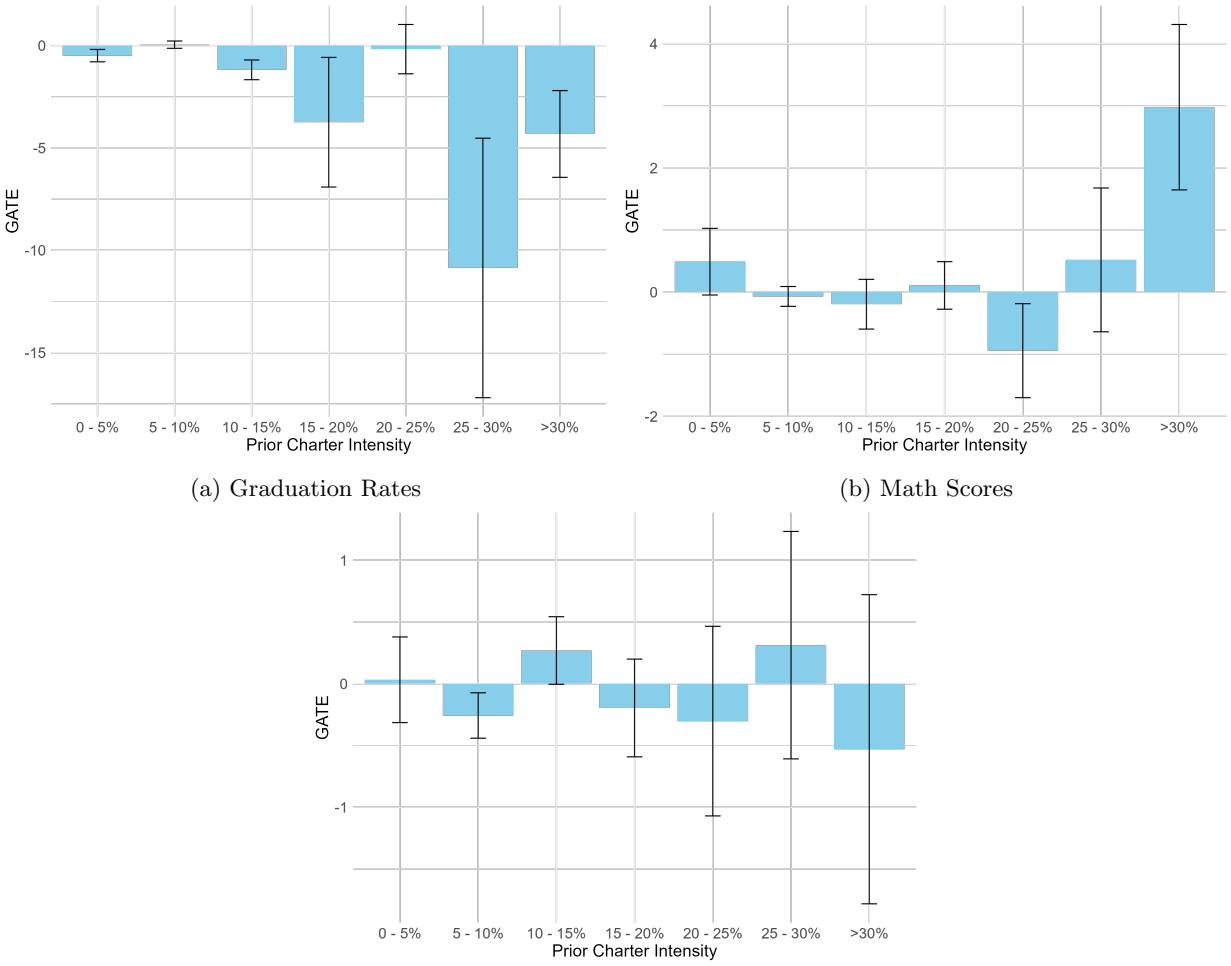
1.4.5 ATEs within States and Dose-Responses

Earlier in the paper, we presented estimates of state-level treatment effects. Table 8 presents the same results as estimated by the causal forest. The causal forest estimates, from what we can tell, are qualitatively quite similar to those produced from interactions in the GDD.

Another measure we presented earlier in the paper was a dose-response: how does the ATE vary by the prior level of treatment? There are two relevant aspects to the dose-response function with continuous treatment:

- The average causal response at different starting doses: $E[Y(d') - Y(d)|\boldsymbol{X}_{it}]$
- The level treatment effect function: $E[Y(d) - Y(0)|\boldsymbol{X}_{it}]$

We can approximately map the average causal response function by grouping units treated in a given period t according to their 1-period lagged dose (i.e. group by $Y(d)_{t-1}$). We group districts by ascending thresholds of this value, then compute GATEs within thresholds, effectively estimating $E[Y(d')_t - Y(d)_{t-1}|\boldsymbol{X}_{i,t-1}]$. Figure 7 plots these results. Like with the GDD, the causal forest estimates do not provide strong evidence for diminishing returns (or any consistent relationship between prior charter intensity and the treatment effect) for math or ELA scores. Though, there is perhaps some evidence for diminishing returns to graduation rates, as GATE estimates become increasingly negative beyond 10% charter share – though this is still a relatively weak relationship.



(a) Graduation Rates

(b) Math Scores

(c) ELA Scores

Figure 7: Dose-Response Treatment Effects

Notes: These figures display average causal responses for increasing 5% thresholds of initial treatment magnitude.

Table 8: GATEs within States

	GATE Estimates			Proportion of N		
State	Grad Rate	Math	ELA	Grad Rate Share	Math Share	ELA Share
Alabama	0.149	NA	NA	0.011	NA	NA
Alaska	-0.009	1.665	-0.073	0.005	0.002	0.002
Arizona	0.016	0.171	0.161	0.010	0.017	0.017
Arkansas	0.103	0.735	-0.207	0.027	0.030	0.030
California	0.031	0.078	0.041	0.040	0.069	0.069
Colorado	-0.305	-0.284**	-0.48***	0.021	0.008	0.008
Connecticut	0.116	0.009	-0.073	0.013	0.017	0.017
Delaware	-0.031	-0.183	-0.124	0.002	0.002	0.002
Florida	0.352***	-0.231	0.034	0.008	0.007	0.007
Georgia	0.261**	2.689***	-0.078	0.020	0.023	0.023
Idaho	0.035	-0.031	0.232	0.012	0.011	0.011
Illinois	0.266**	-0.172	-0.267	0.047	0.082	0.082
Indiana	0.129	-0.234*	-0.006	0.034	0.039	0.039
Kansas	-0.072	-0.359	-0.168	0.033	0.028	0.028
Kentucky	-0.145	-0.653	-0.464	0.017	0.022	0.022
Louisiana	-0.309***	-0.056	0.027	0.008	0.009	0.009
Maine	-0.024	NA	NA	0.012	NA	NA
Massachusetts	0.024	-0.051	0.111	0.026	0.028	0.028
Michigan	-0.124	-0.145	0.018	0.060	0.065	0.065
Minnesota	-0.022	-0.29	-0.153	0.032	0.040	0.040
Mississippi	0.083	-0.235***	0.044	0.012	0.016	0.016
Montana	0.164	0.192	0.304	0.008	0.010	0.010
Nebraska	0.559	-0.013	0.589	0.026	0.018	0.018
Nevada	-0.236*	NA	NA	0.002	NA	NA
New Hampshire	0.047	-0.031	0.039	0.007	0.011	0.011
New Jersey	0.123***	0.598	-1.228	0.021	0.049	0.049
New Mexico	-0.151	0.106	-0.245*	0.010	0.007	0.007
New York	0.089	-0.026	-0.063	0.070	0.042	0.042
North Carolina	0.075	-0.176*	0.074	0.012	0.015	0.015
North Dakota	0.13	-1.178	0.46	0.016	0.007	0.007
Ohio	0.03	0.081	-0.046	0.072	0.063	0.063
Oregon	-0.064	-0.159	-0.035	0.020	0.015	0.015
Pennsylvania	-0.042	-0.067	-0.419**	0.055	0.065	0.065
South Carolina	-0.09	-0.107	-0.103*	0.008	0.011	0.011
South Dakota	0.632	0.141	-0.295	0.017	0.012	0.012
Texas	0.13***	-0.137	-0.042	0.111	0.048	0.048
Utah	0.077	-0.086	0.236	0.004	0.004	0.004
Vermont	-0.145	NA	NA	0.004	NA	NA
Virginia	-0.058	NA	NA	0.015	NA	NA
Washington	0.173**	NA	NA	0.025	NA	NA
West Virginia	1.86	-0.402	-0.9	0.006	0.005	0.005
Wisconsin	0.085*	-0.121	0.026	0.041	0.051	0.051

References

Abadie, A. (2003, April). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231–263.

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

Athey, S. and G. W. Imbens (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics 11*(1), 685–725. _eprint: https://doi.org/10.1146/annurev-economics-080217-053433.

Athey, S., J. Tibshirani, and S. Wager (2019, April). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178. Publisher: Institute of Mathematical Statistics.

Barber, B. M. and J. D. Lyon (1996, July). Detecting abnormal operating performance: The empirical power and specification of test statistics. *Journal of Financial Economics 41*(3), 359–399.

Borusyak, K., X. Jaravel, and J. Spiess (2024, February). Revisiting Event-Study Designs: Robust and Efficient Estimation. *The Review of Economic Studies*, rdae007.

Breiman, L. (2001, October). Random Forests. *Machine Learning 45*(1), 5–32.

Callaway, B. and P. H. C. Sant'Anna (2021, December). Difference-in-Differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

de Chaisemartin, C. and X. D'Haultfœuille (2023, September). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: a survey. *The Econometrics Journal 26*(3), C1–C30.

Gulen, H., C. Jens, and T. B. Page. The heterogeneous effects of default on investment: An application of causal forest in corporate finance. *Working Paper*.

Hirano, K. and J. R. Porter (2009). Asymptotics for Statistical Treatment Rules. *Econometrica 77*(5), 1683–1701. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6630.

Kattenberg, M., B. Scheer, and J. Thiel (2023). Causal forests with fixed effects for treatment effect heterogeneity in difference-in-differences.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica 56*(4), 931–954. Publisher: [Wiley, Econometric Society].

Roth, J. (2022, September). Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. *American Economic Review: Insights 4*(3), 305–322.

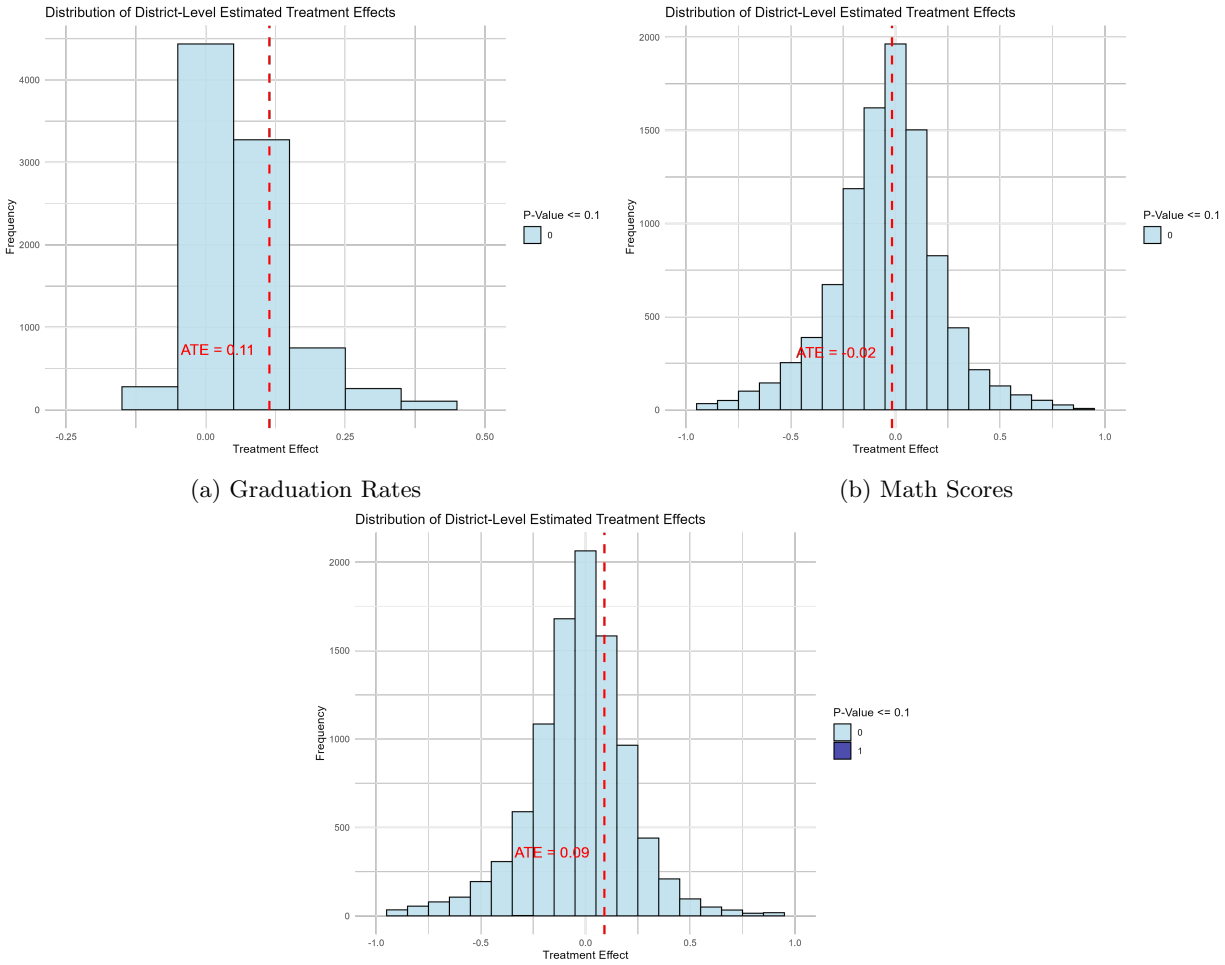
Roth, J., P. H. C. Sant'Anna, A. Bilinski, and J. Poe (2023, August). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics 235*(2), 2218–2244.

Sun, L. and S. Abraham (2021, December). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

Wager, S. and S. Athey (2018, July). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association 113*(523), 1228–1242. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1319839.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

A Appendix A: Additional Figures



(a) Graduation Rates

(b) Math Scores

(c) ELA Scores

Figure 8: Distributions of District-Level ITE Estimates

Notes: These figures are the distributions of individualized treatment effects aggregated to the school district level for each of the three main outcomes from the causal forests. We estimate standard errors for each ITE, and shade darker those districts with statistically significant (at the $p = 0.1$ level) ITE estimates. We also display the GRF-estimated ATE which comes from the aggregation of doubly-robust scores. These estimates are similar to the district $\times$ year estimates, but are aggregated over years to abstract from the time element and obtain a district-level estimate that is comparable to the district-level estimates in the first part of the paper. These are interpreted as average partial effects for a given district. That is, each point represents the predicted treatment effect from increasing the charter share in district d by 1 percentage point.

B Appendix B: Additional Tables

Table 9: Comparing Covariate Means: Graduation Rates

Covariate	All Positive	All Negative	Difference (Positive - Negative)
Log of Enrollment	7.42	7.49	-0.07***
White (%)	0.79	0.82	-0.03***
Black (%)	0.07	0.08	-0.01***
Hispanic (%)	0.09	0.07	0.03***
Free/Reduced Lunch (%)	0.28	0.26	0.03***
Special Ed (%)	0.13	0.12	0.01***
Baseline Performance	0.81	0.80	0.01***
Urban	0.06	0.06	0
Suburb	0.22	0.23	-0.01**
Town	0.19	0.18	0.01***
Rural	0.53	0.53	0
Magnet Schools (%)	0.05	0.02	0.03***
City Population (standardized)	0.00	0.00	0*
University in City	0.00	0.00	0***
Per Pupil Revenue	9504.96	9113.57	391.38***
Student-Teacher Ratio	14.99	15.15	-0.16***
Teacher Salary	72519.19	71398.15	1121.04***
TPS-Charter Spending (% diff)	0.02	0.02	0
Total Spending (per-pupil)	9578.89	9176.46	402.43***
Equitable Funding	5.80	5.26	0.54***
No Caps on CS Growth	8.87	8.97	-0.1***
Performance-Based Contracts	9.21	9.41	-0.2***
Transparent Charter Startup Policies	8.71	9.18	-0.47***
Clear Charter Renewal Policies	10.46	10.59	-0.12***
Exempt from State/District Regs	7.24	7.40	-0.16***
Number of Observations	90065.00	42809.00	132874

Table 10: Comparing Covariate Means: Math

Covariate	Significantly Positive	Significantly Negative	Difference (Positive - Negative)
Log of Enrollment	7.45	7.49	-0.04***
White (%)	0.70	0.71	-0.01***
Black (%)	0.08	0.08	0***
Hispanic (%)	0.14	0.13	0.01***
Free/Reduced Lunch (%)	0.48	0.48	0.01***
Special Ed (%)	0.14	0.14	0***
Baseline Performance	0.10	0.13	-0.04***
Urban	0.06	0.06	-0.01***
Suburb	0.27	0.28	0***
Town	0.19	0.19	0.01***
Rural	0.47	0.47	0***
Magnet Schools (%)	0.17	0.27	-0.1***
City Population (standardized)	0.00	0.00	0***
University in City	0.00	0.00	0***
Per Pupil Revenue	13340.23	13362.44	-22.21**
Student-Teacher Ratio	15.24	15.13	0.11***
Teacher Salary	94666.86	94338.90	327.97***
TPS-Charter Spending (% diff)	0.02	0.02	0.01***
Total Spending (per-pupil)	13213.70	13225.61	-11.91
Equitable Funding	5.63	5.60	0.02**
No Caps on CS Growth	8.84	8.80	0.04***
Performance-Based Contracts	8.62	8.66	-0.03***
Transparent Charter Startup Policies	8.39	8.44	-0.05***
Clear Charter Renewal Policies	10.41	10.45	-0.03***
Exempt from State/District Regs	7.26	7.21	0.04***
Number of Observations	229182.00	248330.00	477512

Table 11: Comparing Covariate Means: ELA

Covariate	Significantly Positive	Significantly Negative	Difference (Positive - Negative)
Log of Enrollment	7.47	7.48	-0.01**
White (%)	0.72	0.70	0.02***
Black (%)	0.08	0.08	0***
Hispanic (%)	0.13	0.14	-0.01***
Free/Reduced Lunch (%)	0.48	0.48	0**
Special Ed (%)	0.14	0.14	0***
Baseline Performance	0.12	0.12	0
Urban	0.06	0.07	-0.01***
Suburb	0.27	0.29	-0.02***
Town	0.20	0.18	0.02***
Rural	0.48	0.47	0.01***
Magnet Schools (%)	0.16	0.29	-0.13***
City Population (standardized)	0.00	0.00	0***
University in City	0.00	0.00	0***
Per Pupil Revenue	13096.46	13600.83	-504.36***
Student-Teacher Ratio	15.28	15.09	0.18***
Teacher Salary	93318.71	95645.00	-2326.29***
TPS-Charter Spending (% diff)	0.02	0.02	0.01***
Total Spending (per-pupil)	12971.90	13461.80	-489.9***
Equitable Funding	5.47	5.75	-0.28***
No Caps on CS Growth	8.83	8.82	0.01
Performance-Based Contracts	8.67	8.61	0.06***
Transparent Charter Startup Policies	8.32	8.50	-0.18***
Clear Charter Renewal Policies	10.40	10.47	-0.07***
Exempt from State/District Regs	7.26	7.21	0.05***
Number of Observations	235790.00	241722.00	477512