Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Using Nonparametric Regression Trees to Estimate Different Forms of Heterogeneous Treatment Effects

G. Buhrman [*†], X. Liao [‡], and J.-S. Kim [†]

†Department of Educational Psychology, University of Wisconsin, Madison, Wisconsin, USA
‡Educational and Counseling Psychology, and Special Education, University of British Columbia, Vancouver, Canada
*Corresponding author. Email: buhrman@wisc.edu

Abstract

Interest in heterogeneous treatment effects has substantially increased in recent years. Treatment heterogeneity describes the case when individuals are differentially affected by an intervention or exposure according to their characteristics, and accurate estimation of these differential effects can support cost-effectiveness evaluations of interventions and inform policy decisions about which individuals or groups will benefit most from an intervention. However, the functional form of heterogeneous treatment effects can vary and is typically unknown to researchers. For instance, the effect of math tutoring on students' test scores might vary across students' prior math scores as a negative quadratic function, meaning that students who benefit most do not have particularly high or low prior scores. Such "Goldilocks" effects and other complex treatment functions have motivated the use of nonparametric regression techniques which make few or no assumptions about the true data generating model. While previous studies have proposed and compared the performance of different nonparametric methods across different datasets, few studies have explicitly explored how the complexity of the functional form of the heterogeneous treatment effects impacts the performance of nonparametric regression tree methods. We initially sought out to explore how the monotonicity of the treatment effect function impacted performance, but present findings that pertain to the overall complexity of the treatment effect function. In these proceedings, we 1) explain why complexity of the treatment effect function is a relevant and important consideration and 2) provide results from a preliminary simulation study which examines how variation in the functional form of treatment effects impacts the accuracy of popular nonparametric regression tree approaches in the context of clustered data with a non-random treatment assignment. We conclude with a discussion of the limitations of the study and possible avenues for future research. Our results suggest that functional complexity, rather than monotonicity, plays a more critical role in the accuracy of nonparametric treatment effect estimators.

Keywords: nonparametric regression, causal inference, heterogeneous treatment effects, Bayesian additive regression trees

1. Introduction

The fields of social and behavioral sciences are in the midst of a heterogeneity revolution. Initiated by a combination of factors including the replicability crisis, disproportionate attention towards main effects, and the lack of attention towards generalizability, this shift in the field of behavioral sciences has motivated research in recent years to increasingly acknowledge and discuss heterogeneous treatment effects (Bryan et al., 2021; Cikara et al., 2022; Hallsworth, 2023; Szaszi et al., 2022; Walton et al., 2023; Yeager et al., 2022). Beyond the revolution, heterogeneous treatment effects are also important to consider when evaluating the cost-effectiveness of interventions or when making policy decisions about which units should receive an intervention.

Heterogeneous treatment effects can be represented as a function of the observed characteristics of individuals and groups in the sample, but this functional form can take a variety of shapes, and the functional form is usually unknown to researchers. In order to best identify who will benefit most from an intervention, and under what conditions, accurate estimation of this functional form is necessary. Previous research has shown that these functional forms can be accurately estimated using nonparametric and machine learning-based methods, but there has been little attention given to how different approaches' results might vary depending on the functional form of the heterogeneous treatment effect. In these proceedings, we attempt to shed light on this area of research by demonstrating the situations where different nonparametric methods return similar or different results depending on the functional form of the heterogeneous treatment effect. We do this by conducting a simulation study using semi-synthetic data generation. to generate clustered data with a non-random treatment assignment.

1.1 Potential Outcomes Framework in Clustered Data

We used the Neyman-Rubin potential outcomes framework for causal inference in this study (Rubin, 1974; Splawa-Neyman et al., 1923). We use the extended notation of potential outcomes for the multilevel structure where units are nested within clusters (Hong & Raudenbush, 2006; Lyu et al., 2023). Let us assume that there are N individuals nested within M clusters. In this scenario, let $Y_{ij}(1)$ denote the potential outcome if individual i within cluster j received treatment ($T_{ij} = 1$) and $Y_{ij}(0)$ denote the potential outcome if individual i in cluster j did not receive treatment ($T_{ij} = 0$), where $i = 1, ..., n_j$ in cluster $j = 1, ..., M$ and $\sum_{j=1}^{M} n_j = N$. Under this framework, the observed outcome can be expressed as

$$Y_{ij} = T_{ij}Y_{ij}(1) + (1 - T_{ij})Y_{ij}(0), \tag{1}$$

under the *stable unit treatment value assumption*, or SUTVA (Rubin, 1986). SUTVA states that the potential outcomes of individuals are not affected by the treatment assignments of other individuals, and that there are no hidden versions of the treatment. Hong and Raudenbush, and Imbens and Rubin, provide more detail about SUTVA in multilevel settings (Hong & Raudenbush, 2006; Hong & Raudenbush, 2013; Imbens & Rubin, 2015). The two potential outcomes $Y_{ij}(1)$ and $Y_{ij}(0)$ can never be observed at the same time for the same individual, meaning that individual treatment effects cannot be calculated outright. However, under certain key assumptions we can express the average treatment effect τ as

$$\tau = \mathrm{E}[Y_{ij}(1) - Y_{ij}(0)]. \tag{2}$$

The first assumption is that potential outcomes are independent of treatment assignment T_{ij}. This can be achieved via random treatment assignment or by establishing unconfoundedness if treatment assignment was non-random. Often referred to as *(conditional) ignorability* (Rosenbaum & Rubin, 1983; Rubin, 1978), this assumption states that potential outcomes are independent of treatment assignment, conditional on individual and cluster covariates, $\mathbf{X_{ij}}$ and $\mathbf{Z_j}$, respectively:

$$\text{Unconfoundedness: } Y_{ij}(1), Y_{ij}(0) \perp T_{ij} \mid \mathbf{X_{ij}}, \mathbf{Z_j} \tag{3}$$

The second assumption necessary for valid causal inference is that the probability that an individual is assigned to either treatment condition, given their underlying characteristics or covariates, is strictly between 0 and 1:

$$\text{Positivity: } 0 < e(\mathbf{X_{ij}}, \mathbf{Z_j}) = \Pr(T_{ij} = 1 \mid \mathbf{X_{ij}}, \mathbf{Z_j}) < 1, \tag{4}$$

where $e(\mathbf{X_{ij}}, \mathbf{Z_j})$ is the *propensity score* (Rosenbaum & Rubin, 1983).

1.2 Heterogeneous Treatment Effects and CATE Estimation

Conditional average treatment effects, or CATEs, are a type of treatment effect that can be estimated to quantify treatment heterogeneity (Imbens & Rubin, 2015). If we believe that a treatment effect may vary across the observed individual-level ($\mathbf{X_{ij}}$) and cluster-level ($\mathbf{Z_j}$) covariates, then the CATEs can be estimated as

$$\tau_{ij} = \mathrm{E}[Y_{ij}(1) - Y_{ij}(0) \mid \mathbf{X_{ij}}, \mathbf{Z_j}]. \tag{5}$$

If we could estimate the treatment effect for each individual i within cluster j, conditional upon the individual-level and cluster-level covariates we would have a vector of individual treatment effects (ITEs), denoted as τ_{ij}, conditional upon each individual's characteristics. These ITEs can be analyzed and visually inspected to determine whether there are heterogeneous treatment effects.

1.3 Motivating Problem

In a previous study by Kim, Liao, and Loh (Kim et al., 2024), the authors estimated the ITEs for students in the 2015 Korea TIMSS data conditional upon a variety of individual-level and cluster-level covariates. Notably, they found evidence of a cross-level interaction between private tutoring and school resource shortages for math instruction (see Figure 1). The authors also found evidence of a second cross-level interaction between private tutoring and the school's emphasis on academic success (see Figure 2).

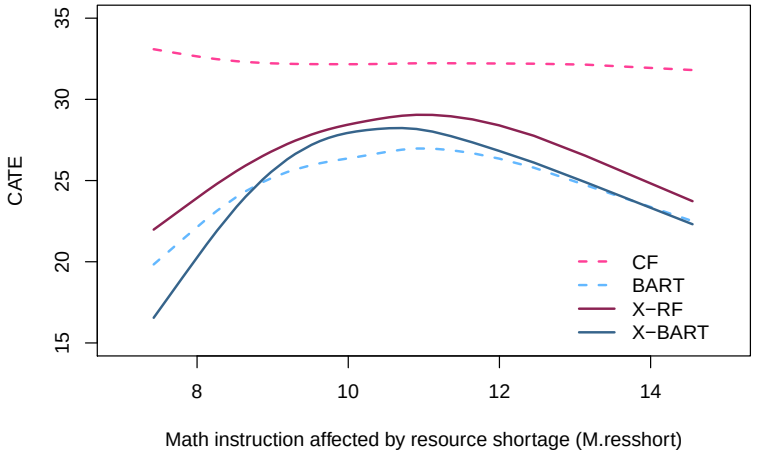

Figure 1. CATE estimates of the impact of private tutoring on students' TIMSS mathematics scores with respect to school-level resource shortage.

Two important observations from these findings are that 1) the functional form of the heterogeneous treatment effect varies (in Figure 1, the form seems to be quadratic, while in Figure 2 the form seems to be linear), and 2) depending on the method you used (CF: Causal Forest; BART: Bayesian Additive Regression Trees; X-RF: X-Learner with Random Forests; X-BART: X-Learner with BART), the amount of heterogeneity you would ascribe to these interactions would vary. These observations are the motivation for the current study. Specifically, we wanted to explore how variation in the functional form of treatment effects impacts the accuracy of popular nonparametric regression tree approaches in the context of clustered data with a non-random treatment assignment. In the following sections, we provide a brief overview of how CATEs can be estimated with nonparametric regression and detail the results from a preliminary simulation study.

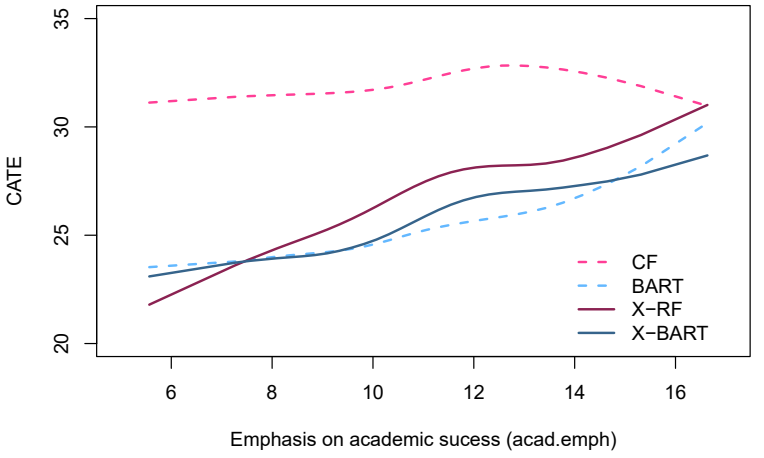

Figure 2. CATE estimates of the impact of private tutoring on students' TIMSS mathematics scores with respect to school-level emphasis on academic success.

2. CATE Estimation via Nonparametric Regression

Without pre-existing knowledge of subgroups and the functional form of heterogeneous treatment effects, CATEs can become difficult to estimate, especially in data with many covariates where interactions between covariates and treatment are potentially numerous and complex.

Nonparametric regression methods, specifically nonparametric regression tree-based methods, are a useful approach for estimating CATEs because they are "agnostic" to the functional form of heterogeneous treatment effects and they can consider a large number of covariates as potential characteristics upon which treatment may vary. In general, nonparametric regression tree-based methods start with the supposition that the observed outcomes can be defined as the output from some "unknown" function $f(\cdot)$:

$$Y_{ij} = f(\mathbf{X_{ij}}, \mathbf{Z_j}, T_{ij}) + \epsilon_{ij}, \tag{6}$$

where $\mathbf{X_{ij}}$ is a matrix of individual-level covariates, $\mathbf{Z_j}$ is a matrix of cluster-level covariates, T_{ij} is an $N \times 1$ column vector of the binary treatment assignment, and ϵ_{ij} is some error term with $\mathrm{E}[\epsilon_{ij}] = 0$ and no distributional assumptions. Nonparametric regression tree-based methods vary from each other in a variety of ways, but in general, their distinguishing characteristics can be organized into four broad domains:

1. The components included in the functional definition of the outcome.
2. CATE estimation as "counterfactual prediction" vs. "effect estimation"
3. The type and targets of regularization.
4. The statistical framework (Bayesian vs. Frequentist)

The first domain refers to variations of the expression in equation 4. For instance, some methods may include cluster identification or ID as an additional input in $f(\cdot)$ resulting in

$$Y_{ij} = f(\mathbf{X_{ij}}, \mathbf{Z_j}, T_{ij}, j) + \epsilon_{ij}, \tag{7}$$

where j now operates in a manner akin to specifying cluster ID as a fixed effect in a parametric regression model.

The second domain refers broadly to the "learner" or approach taken to estimate CATEs (Caron et al., 2022a; Künzel et al., 2019). Tran and colleagues coined the terms "counterfactual prediction" and "effect estimation" as ways to refer to the most common approaches for estimating CATEs with nonparametric methods (Tran et al., 2024). Counterfactual prediction refers to approaches that estimate two models, $f_1(\mathbf{X_{ij}}, \mathbf{Z_j}, T_{ij} = 1)$ and $f_0(\mathbf{X_{ij}}, \mathbf{Z_j}, T_{ij} = 0)$, for treated and control units respectively. These fitted models can then be used to predict the unobserved potential outcome for each unit. Taking the difference between the observed and predicted potential outcomes gives us an estimate of the ITE conditioned upon the individual-level and cluster-level covariates. Examples of methods that use counterfactual prediction are BART as a T-learner (Hill, 2011) and stan4bart (S4BART) (Dorie et al., 2022).

Effect estimation refers to approaches that directly estimate the CATEs by re-specifying the formula in equation 6 to take the general form of

$$Y_{ij} = f(\mathbf{X_{ij}}, \mathbf{Z_j}, T_{ij}) + \epsilon_{ij} = \mu(\mathbf{X_{ij}}, \mathbf{Z_j}) + \tau(\mathbf{X_{ij}}, \mathbf{Z_j}) \times T_{ij} + \epsilon_{ij}, \tag{8}$$

where $\mu(\mathbf{X_{ij}}, \mathbf{Z_j})$ is a function that gives the prognostic outcome, and $\tau(\mathbf{X_{ij}}, \mathbf{Z_j})$ is a function that gives the individual's CATE. The function $\tau(\mathbf{X_{ij}}, \mathbf{Z_j})$ can be estimated using nonparametric regression tree-based approaches, such as BART. Examples of methods that use effect estimation are Bayesian Causal Forest (BCF) (Hahn et al., 2020) and CF (Wager & Athey, 2018).

The third domain refers to which parts of the estimation algorithm or function include regularization and what type of regularization procedure is utilized. In BART, overfitting is mitigated by nature of the approach being an "ensemble-method" where many "weak-learners" are combined to provide a full picture of the data. BART is able to combine many small regression trees, and these trees are kept shallow, meaning they have a few numbers of cut/decision points, via a Bayesian prior. In this situation, the target of regularization are the regression trees, and the type of regularization is a Bayesian shrinkage prior.

The final domain, the statistical framework chosen, is self-explanatory, but also has implications for the third domain, as working within a certain framework gives access to different types of regularization procedures. Namely, that working in a Bayesian framework allows for the use of Bayesian shrinkage priors for regularization. In the current study, we identified methods that were both popular and varied across these four domains in meaningful ways. Namely, we consider CF (Wager & Athey, 2018), BCF (Hahn et al., 2020), Sparse Bayesian Causal Forest (SBCF) (Caron et al., 2022b), and S4BART (Dorie et al., 2022). The differences between these selected methods are summarized in Table 1 and Table 2. For further details about each of these approaches, see their respective references.

Table 1. Selected methods' estimation and regularization approaches

Method	CATE Estimation	Regularization Types	Regularization Targets
BCF-FE	Effect Estimation	BART prior	CATE, prognostic outcome, trees
SBCF-FE	Effect Estimation	BART prior & Dirichlet Prior	CATE, prognostic outcome, splitting, trees
CF-FE	Effect Estimation	Adaptive kernel weighting	Nuisance functions, CATE
S4BART	Counterfactual Prediction	BART prior	Trees

3. Simulation: Semi-Synthetic TIMSS data

We conducted a simulation study to answer the following questions: (1) for what functional forms of heterogeneous treatment effects will different nonparametric methods agree and (2) for what forms of heterogeneous treatment effects will different nonparametric methods disagree? The data contexts in which we are interested are observational studies with clustered data, so we wanted to

Table 2. Selected methods' outcome specifications

Method	Outcome Specification
BCF-FE	$Y_{ij} = \mu(\mathbf{X_{ij}}, \mathbf{Z_j}, j) + \tau(\mathbf{X_{ij}}, \mathbf{Z_j}, j) \times T_{ij} + \epsilon_{ij}$
SBCF-FE	$Y_{ij} = \mu(\mathbf{X_{ij}}, \mathbf{Z_j}, j) + \tau(\mathbf{X_{ij}}, \mathbf{Z_j}, j) \times T_{ij} + \epsilon_{ij}$
CF-FE	$Y_{ij} = (T_{ij} - \pi(\mathbf{X_{ij}}, \mathbf{Z_j}, j))\tau(\mathbf{X_{ij}}, \mathbf{Z_j}, j) - m(\mathbf{X_{ij}}, \mathbf{Z_j}, j) + \epsilon_{ij}$
S4BART	$Y_{ij} = f(\mathbf{X_{ij}}, \mathbf{Z_j}, T_{ij}) + U_{0j} + \epsilon_{ij}$

use a data generating procedure that would mimic these situations. To accomplish this, we used a semi-synthetic data generation process (Buhrman et al., 2024; Hill, 2011). Our semi-synthetic approach differs from previous approaches in that we generated covariates based on the covariance structure of real data rather than randomize or sample from real data. Specifically, we used covariates from the 2019 United States TIMSS data to obtain the covariance structure used in our simulation. All analyses were performed using R Statistical Software (R Core Team, 2025). Implementation of BCF and SBCF was performed using the SparseBCF package (Caron, 2020), implementation of CF was performed using the grf package (Tibshirani et al., 2024), and implementation of S4BART was performed using the stan4bart package (Dorie, 2024).

3.1 Data and Variables

The 2019 United States TIMSS data includes several context variables for both students and schools. Student and family-related covariates we used to generate a covariance structure for the data generation process include student gender, household socioeconomic status, student's confidence in math, student's fondness for math, student's value of math, and the number of absences a student had. School- or cluster-level covariates include the percentage of male students, the average SES of students, the school's emphasis on academic success, the school's strictness in disciplinary policies, and the degree to which mathematics instruction was affected by school resource shortage. We school-mean centered student-level covariates prior to obtaining the covariance structures for student- and school-level covariates. Using these covariance structures, we generated clustered data with random intercepts.

We generated data for 30 clusters with cluster size ranging from 22 to 38 with an average cluster size of 30. The unconditional interclass correlation (ICC) was 0.15. Our goal was to generate data for the effect of some non-random individual-level treatment assignment on student's math performance, where this treatment effect followed different functional forms according to some other covariate. We generated an individual-level non-random treatment assignment, which we framed as student participation in extra-curricular math activities like Math Olympiad, based on the following propensity score function

$$\pi_{ij} = \text{logit}(-0.25 - 0.25Absences_{ij} + 0.25SES_{ij} + 0.5Confidence_{ij} + 0.25Emphasis_j - 0.75Shortage_j + W_j), \quad (9)$$

where π_{ij} is the probability that student i in school j participates in an extra-curricular math activity, $W_j \sim N(0, 0.01)$ and the binary treatment indicator is $T_{ij} \sim \text{Bernoulli}(\pi_{ij})$.

We also generated heterogeneous treatment effects based on three different functional forms: (1) linear, (2) quadratic, and (3) logistic. Each form can be thought of as a cross-level interaction between a school-level covariate and the treatment indicator. These can be expressed as

$$\tau_{ij(linear)} = \frac{Shortage_j + 2}{8}, \quad (10)$$

$$\tau_{ij(quadratic)} = \frac{-Shortage_j^2 + 5}{10}, \quad (11)$$

$$\tau_{ij(logistic)} = \frac{0.5}{1 + e^{-Shortage_j/0.25}}. \tag{12}$$

We generated 1000 iterations of data under each of these functional form conditions and estimated the ITEs conditional on covariates for each iteration using each of the four methods previously described. This left us with the true ITEs conditional on covariates and the estimated ITEs conditional on covariates for 1000 iterations of each functional form specification. As an aside, we estimated propensity scores the same way for every method, using BART with random intercepts.

To quantify the performance of each method, we used the Precision of Estimation of Heterogeneous Effects (PEHE) (Hill, 2011) to evaluate and compare methods' performances. The PEHE is a measure of both bias and variance of the estimated ITEs conditioned on covariates. However, because our simulated data are observational, it is possible for there to be regions of non-overlap for certain covariates. Estimating CATE on regions of non-overlap produces biased estimates, so we use PEHE on treated, or PEHET, as the evaluation criteria. PEHET can be expressed as

$$\text{PEHET} = \sqrt{\frac{1}{N_T}\sum_{i}^{N_T}(\tau_i - \hat{\tau}_i)^2}, \tag{13}$$

where N_T is the number of treated units, τ_i is the true ITE conditional on covariates for individual i, and $\hat{\tau}_i$ is the estimated ITE conditional on covariates for individual i.

3.2 Results

We find that stan4bart consistently recovers the true heterogeneous treatment effect with greater accuracy and less variance compared to the other three methods (see Table 3 and Figure 3). We also observe that all the methods were most accurate when the heterogeneous treatment effect took a linear form, and least accurate when the treatment effect took a logistic form (see Figure 3). This was especially the case for Causal Forest, for which we observe the distribution of PEHETs under the logistic form condition is highly separated from the distributions of PEHETs under the linear and quadratic form conditions. Recalling our initial interest in determining whether the monotonicity of the functional form mattered for heterogeneous treatment estimation, it initially looks like monotonicity does matter, until you consider the complexity of the functional form. If you were to only compare linear and quadratic forms, you would come to the conclusion that it is more difficult for nonparametric regression tree methods to estimate non-monotonic functional forms. However, when you consider the logistic form, which is monotonic, we can see that it is functional complexity, not monotonicity, driving the pattern observed in the simulation results.

Table 3. PEHET statistics for each method and functional form condition across 1000 iterations

	Linear Form		Quadratic Form		Logistic Form	
	Mean (SD)	Median [Min, Max]	Mean (SD)	Median [Min, Max]	Mean (SD)	Median [Min, Max]
BCF-FE	0.147 (0.040)	0.140 [0.076, 0.313]	0.184 (0.044)	0.177 [0.082, 0.423]	0.205 (0.040)	0.203 [0.096, 0.364]
SBCF-FE	0.165 (0.044)	0.161 [0.051, 0.366]	0.202 (0.047)	0.198 [0.057, 0.410]	0.226 (0.054)	0.232 [0.063, 0.396]
CF-FE	0.158 (0.031)	0.152 [0.086, 0.317]	0.175 (0.046)	0.169 [0.084, 0.413]	0.238 (0.043)	0.242 [0.084, 0.389]
S4BART	0.135 (0.040)	0.130 [0.051, 0.313]	0.156 (0.042)	0.149 [0.066, 0.360]	0.153 (0.045)	0.150 [0.051, 0.298]

4. Discussion

Estimation of heterogeneous treatment effects has become an increasingly important topic of research, especially in clustered data where the contexts of membership in different clusters may change the

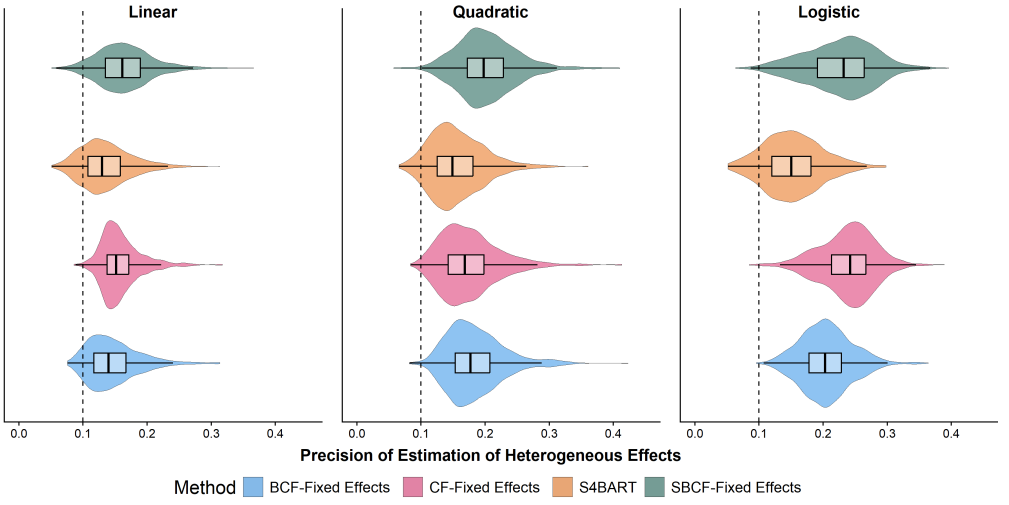

Figure 3. Distributions of PEHETs by method for each functional form condition

degree to which individuals benefit from an intervention. Nonparametric regression trees are a popular technique for estimating heterogeneous treatment effects because of their flexibility and usability. However, little research has paid attention to how the functional form of the treatment effect affects the performance of specific nonparametric methods and the family of methods overall. In a preliminary study, we initially hypothesized that monotonicity of the functional form may impact the accuracy and variance of ITE estimation, but found evidence that functional complexity was the driving factor for differences across all the methods we investigated. This chapter details these findings and outlines the key differences between the methods we investigated.

The goal of this work is to find conditions in which the results of different methods are consistent and to compare these to conditions where the results from different methods are inconsistent. By finding these conditions, we can begin to identify the characteristics of methods which might be most relevant for certain conditions. We can apply these findings to practice in the form of method selection when the objective is to estimate heterogeneous treatment effects in plausible scenarios, including observational data with uneven treatment allocation, data with multiple sources of treatment heterogeneity, and clustered data with small or varying cluster sizes.

For the findings presented in this chapter, we suspect that the differences in the accuracy and variance of ITE estimation can be explained with the four domains we outlined in section 2. Specifically, we hypothesize that the types and targets of regularization and the way that CATE is estimated play the largest role in observed differences depending on the condition of functional form. Based on the results of this study, we recommend the use of stan4bart for general purpose estimation of heterogeneous treatment effects from multilevel data when treatment is assigned at the lowest level. However, stan4bart makes a distributional assumption to model the variance between clusters, meaning that users should still perform diagnostic checks on the parametric component of the model. Future work in this area could consider the four domains we have described and should consider functional forms that include multiple covariates. Furthermore, future research should continue to explore how nonparametric regression tree methods can be best applied in the context of multilevel data.

Acknowledgment

The authors thank proceedings editor Okan Bulut for his helpful comments.

Funding Statement The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Award #R305B200026 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Competing Interests None.

References

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Buhrman, G., Liao, X., & Kim, J.-S. (2024). Exploring conceptual differences among nonparametric estimators of treatment heterogeneity in the context of clustered data. In M. Wiberg, J.-S. Kim, H. Hwang, H. Wu, & T. Sweet (Eds.), *Quantitative Psychology* (pp. 261–274). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-55548-0_25

Caron, A. (2020). *Sparsebcf: Sparse Bayesian causal forest for heterogeneous treatment effects estimation* [R package version 1.0, commit c7a8b66e812c7df0553d635decde929f09eeaf1f]. https://github.com/albicaron/SparseBCF

Caron, A., Baio, G., & Manolopoulou, I. (2022a). Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 1115–1149. https://doi.org/10.1111/rssa.12824

Caron, A., Baio, G., & Manolopoulou, I. (2022b). Shrinkage Bayesian causal forests for heterogeneous treatment effects estimation. *Journal of Compuational and Graphical Statistics*, *31*(4), 1202–1214. https://doi.org/10.1080/10618600.2022.2067549

Cikara, M., Fouka, V., & Tabellini, M. (2022). Hate crime towards minoritized groups increases as they increase in sized-based rank. *Nature Human Behaviour*, *6*(11), 1537–1544. https://doi.org/10.1038/s41562-022-01416-5

Dorie, V. (2024). *Stan4bart: Bayesian additive regression trees with Stan-sampled parametric extensions* [R package version 0.0-10]. https://CRAN.R-project.org/package=stan4bart

Dorie, V., Perrett, G., Hill, J. L., & Goodrich, B. (2022). Stan and BART for causal inference: Estimating heterogeneous treatment effects using the power of Stan and the flexibility of machine learning. *Entropy*, *24*(12), 1782. https://doi.org/10.3390/e24121782

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, *15*(3), 965–1056. https://doi.org/10.1214/19-BA1195

Hallsworth, M. (2023). A manifesto for applying behavioural science. *Nature Human Behaviour*, 7(3), 310–322. https://doi.org/10.1038/s41562-023-01555-3

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, *101*(475), 901–910. https://doi.org/10.1198/016214506000000447

Hong, G., & Raudenbush, S. W. (2013). Heterogeneous agents, social interactions, and causal inference. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 331–352). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_16

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kim, J.-S., Liao, X., & Loh, W. W. (2024). Assessing cross-level interactions in clustered data using cate estimation methods. In M. Wiberg, J.-S. Kim, H. Hwang, H. Wu, & T. Sweet (Eds.), *Quantitative Psychology* (pp. 87–97). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-55548-0_9

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Lyu, W., Kim, J.-S., & Suk, Y. (2023). Estimating heterogeneous treatment effects within latent class multilevel models: A Bayesian approach. *Journal of Educational and Behavioral Statistics*, *48*(1), 3–36. https://doi.org/10.3102/10769986221115446

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*(1), 34–58. https://doi.org/10.1214/aos/1176344064

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*(396), 961–962. https://doi.org/10.1080/01621459.1986.10478355

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. P. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, *5*(4), 465–472.

Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S., & Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences*, *119*(31), e2200732119. https://doi.org/10.1073/pnas.2200732119

Tibshirani, J., Athey, S., Sverdrup, E., & Wager, S. (2024). *Grf: Generalized random forests* [R package version 2.4.0]. https://CRAN.R-project.org/package=grf

Tran, C., Burghardt, K., Lerman, K., & Zheleva, E. (2024). Data-driven estimation of heterogeneous treatment effects. https://doi.org/10.48550/arXiv.2301.06615

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

Walton, G. M., Murphy, M. C., Logel, C., Yeager, D. S., Goyer, J. P., Brady, S. T., Emerson, K. T. U., Paunesku, D., Fotuhi, O., Blodorn, A., Boucher, K. L., Carter, E. R., Gopalan, M., Henderson, A., Kroeper, K. M., Murdock-Perriera, L. A., Reeves, S. L., Ablorh, T. T., Ansari, S., … Krol, N. (2023). Where and with whom does a brief social-belonging intervention promote progress in college? *Science*, *380*(6644), 499–505. https://doi.org/10.1126/science.ade4420

Yeager, D. S., Carroll, J. M., Buontempo, J., Cimpian, A., Woody, S., Crosnoe, R., Muller, C., Murray, J., Mhatre, P., Kersting, N., Hulleman, C., Kudym, M., Murphy, M., Duckworth, A. L., Walton, G. M., & Dweck, C. S. (2022). Teacher mindsets help explain where a growth–mindset intervention does and doesn't work. *Psychological Science*, *33*(1), 18–32. https://doi.org/10.1177/09567976211028984