Journal of Data Science 21 (2), 391–411
April 2023
DOI: 10.6339/23-JDS1084
Statistical Data Science

Random Forest of Interaction Trees for Estimating Individualized Treatment Regimes with Ordered Treatment Levels in Observational Studies

Justin Thorp[1], Richard A. Levine[1,2,*], Luo Li[1], and Juanjuan Fan[1,2,*]

[1]*Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA*

[2]*Analytic Studies & Institutional Research, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA*

Abstract

Traditional methods for evaluating a potential treatment have focused on the average treatment effect. However, there exist situations where individuals can experience significantly heterogeneous responses to a treatment. In these situations, one needs to account for the differences among individuals when estimating the treatment effect. Li et al. (2022) proposed a method based on random forest of interaction trees (RFIT) for a binary or categorical treatment variable, while incorporating the propensity score in the construction of random forest. Motivated by the need to evaluate the effect of tutoring sessions at a Math and Stat Learning Center (MSLC), we extend their approach to an ordinal treatment variable. Our approach improves upon RFIT for multiple treatments by incorporating the ordered structure of the treatment variable into the tree growing process. To illustrate the effectiveness of our proposed method, we conduct simulation studies where the results show that our proposed method has a lower mean squared error and higher optimal treatment classification, and is able to identify the most important variables that impact the treatment effect. We then apply the proposed method to estimate how the number of visits to the MSLC impacts an individual student's probability of passing an introductory statistics course. Our results show that every student is recommended to go to the MSLC at least once and some can drastically improve their chance of passing the course by going the optimal number of times suggested by our analysis.

Keywords *educational data mining; generalized propensity scores; individualized treatment effect; machine learning; student success study*

1 Introduction

When evaluating the effectiveness of a proposed treatment, traditional methods have focused on estimating the treatment effect for the population average. Issues arise when subjects can have drastically different responses to the proposed treatment, where some subjects may receive a very positive treatment effect while others may experience no or even negative effects. In these situations one needs to account for the differences in subjects' characteristics when evaluating treatment effects for individual subjects. In other words, instead of estimating the average treatment effect (ATE), we may want to estimate the individualized treatment effect

*Corresponding author. Email: rlevine@sdsu.edu or jjfan@sdsu.edu.

Received July 31, 2022; Accepted January 6, 2023

(ITE). By doing so we can recommend better treatment regimens to individual subjects and also potentially reduce cost.

In a randomized controlled trial (RCT), treatment is randomly assigned and then the treatment and control groups are compared to each other to estimate the treatment effect. The randomization ensures that the treatment and control groups are similar with respect to confounding and other variables. However, there exist situations where a RCT is either impractical or unethical. For example, when the treatment being evaluated can cause severe harm to an individual. In such situations we have to rely on observational study data where treatments are not randomly assigned. However, allowing individuals to select their own treatment introduces bias in the estimated treatment effect. For example, a prevalent source of bias in observational study data is that of self-selection, where characteristics that impact an individual's outcome may also impact their likelihood of choosing specific treatment options. One way to control for selection bias is to use the propensity score method (Rosenbaum and Rubin, 1983). Propensity score is defined as the probability of a subject being in the treatment group given their individual characteristics. Rosenbaum and Rubin (1983) show that treatment assignment and covariate values are independent, conditional on the propensity score. This eliminates selection bias in observational studies by balancing the covariates between treatment and control groups, so unbiased estimates of the treatment effect can be attained. In a RCT the propensity score for each individual is known (equal to one-half for everyone in a study with two treatment arms), while in observational studies the propensity score must be estimated from the data. The most common methods used to estimate propensity scores include logistic regression, and more recently machine learning classification methods such as random forest (Breiman, 2001).

Because tree based methods (Breiman et al., 1984) are non-parametric in nature and can handle complex data structures such as variable interactions, they have become a popular and effective approach for estimating ITE. Examples of such methods include random forest of interaction trees (RFIT) (Su et al., 2018) where the splitting rule is modified to maximize subgroup treatment heterogeneity, qualitative interaction trees (Dusseldorp and Mechelen, 2013) where the terminal nodes are split based on the sign and magnitude of the treatment effect, causal random forest (Wager and Athey, 2018) where each terminal node is treated as its own randomized experiment, and subgroup identification based on differential effect search (Lipkovich et al., 2011) where the optimal split creates a subgroup with the largest positive treatment effect. Tree based methods can be thought of as a subgroup identification algorithm, so these methods excel when there exist subgroups that experience drastically different responses to the treatment under study.

Existing methods for estimating ITE and learning individualized treatment regimes (ITR) are usually designed for binary treatment options only. Li et al. (2022) extends RFIT to accommodate observational data with multiple treatments. We propose a further modification to RFIT in order to estimate ITE and ITR when the treatment is an ordered variable. This paper is motivated by an educational data mining problem: evaluating the effect of a university tutoring center on performance in STEM courses. The binary "treatment" may be whether a student attends the tutoring center at least once during the semester. Of course, such a binary treatment ignores dosage, namely how often a student goes to the tutoring center throughout the semester. To handle this ordered treatment, we propose taking advantage of the randomization scheme in random forest by randomly dichotomizing treatment into a binary variable at each split point. This modification improves upon Li et al. (2022) for multiple treatments since the entire data set is considered at each node, which allows more informative trees to be grown while reducing the total number of trees required to achieve a respectable estimate. In addition, preserving the

ordered structure of the treatment variable allows us to more accurately estimate ITE across all levels of treatment.

The remainder of the paper is structured as follows. In Section 2, we first review RFIT (Su et al., 2018), generalized propensity scores, and causal effect random forest of interaction trees for multiple treatments (Li et al., 2022). We then present our proposed method for estimating ITE and ITR based on observational study data with ordered treatments. In Section 3, we present a simulation study comparing our proposed method to competing methods through the metrics of mean squared error for the predicted response, optimal treatment classification, and accuracy of variable importance rankings. In Section 4, we apply our proposed method to student success data, estimating the effect of a student's visit to a math and stat tutoring center on the probability of passing an introductory statistics course at San Diego State University (SDSU). In Section 5, we conclude the paper with a brief discussion.

2 Methods

2.1 Random Forest of Interaction Trees

RFIT was designed to estimate the ITE for binary treatments, namely a treated group and a control group. The challenge is that we require an individual's responses under treatment and control, but in reality each subject is observed under either treatment or control, not both. To solve this problem, RFIT uses Rubin's potential outcome model (Rubin, 1974) where it estimates the response under treatment for control subjects and that under no treatment for treated subjects. We define $Y_i(0)$ as the response for the ith individual when not given treatment and $Y_i(1)$ as the response for the ith individual when given treatment. RFIT estimates these quantities and uses them to estimate the ITE as $\widehat{ITE}_i = \hat{Y}_i(1) - \hat{Y}_i(0)$.

RFIT is an extension of the random forest (RF) machine learning method (Breiman, 2001) for the purpose of estimating the ITE using RCT data. RFIT differs from the standard RF in that each tree grown in the forest is an interaction tree as defined in Su et al. (2009). In particular, the splitting rule at each node is sought to maximize subgroup treatment heterogeneity instead of maximizing the difference in response between the two child nodes. Let t denote a node under consideration for a potential split, and t_L and t_R the left and right child nodes of t. Let $Y_L(1)$ and $Y_L(0)$ represent collections of responses in the left child node for treated and control subjects, respectively. The sets $Y_R(1)$ and $Y_R(0)$ are defined similarly for the right child node.

	t_L	t_R
$Z = 1$	$Y_L(1)$, n_1, s_1^2	$Y_R(1)$, n_3, s_3^2
$Z = 0$	$Y_L(0)$, n_2, s_2^2	$Y_R(0)$, n_4, s_4^2

Here $Z \in \{0, 1\}$ denotes treatment assignment with 1 and 0 indicating treated and control groups respectively, $\{n_1, n_2, n_3, n_4\}$ denote the sample sizes, and $\{s_1^2, s_2^2, s_3^2, s_4^2\}$ denote the sample variances, in each cell of the 2×2 table, cross tabulated by the L/R child nodes and treatment assignment. The split that is chosen is the one that maximizes the difference in ATE between the two child nodes. Let $ATE_L = \frac{1}{n_1}\Sigma Y_L(1) - \frac{1}{n_2}\Sigma Y_L(0)$ and $ATE_R = \frac{1}{n_3}\Sigma Y_R(1) - \frac{1}{n_4}\Sigma Y_R(0)$ denote the ATE in the left and right child nodes, respectively. Here $\Sigma Y_L(1)$, $\Sigma Y_L(0)$, $\Sigma Y_R(1)$, and $\Sigma Y_R(0)$ denote the sum of responses for all the treated or control observations in the left or right child nodes. For example, $\Sigma Y_L(1) = \Sigma_i Y_i \cdot I\{i\text{th observation is sent to the left child node } \& \ Z_i = 1\}$, where $I\{\cdot\}$ is the indicator function, gives the sum of responses for all the treated observations

in the left child node. To measure the difference in ATE_L and ATE_R, the following test statistic is used:

$$\frac{ATE_L - ATE_R}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}+\frac{1}{n_4}}}, \tag{1}$$

where $\sigma = \sum_{i=1}^{4} \frac{(n_i-1)s_i^2}{n-4}$ and $n = \sum_{i=1}^{4} n_i$. It can be shown that the above test statistic is equivalent to the t-test statistic for β_3 in the model

$$Y_i = \beta_0 + \beta_1 I(Z_i = 1) + \beta_2 I(X_{ij} < c) + \beta_3 I(Z_i = 1) I(X_{ij} < c) + \varepsilon_i. \tag{2}$$

Here Y_i denotes the ith response, Z_i denotes the treatment for the ith subject, X_{ij} denotes the splitting variable, and c denotes the potential cut-point. The splitting variable and the corresponding cut-point c that maximizes the squared value of the test statistic for β_3 is chosen as the best split. Each tree is grown using a bootstrap sample and only a subset of variables, of size $mtry$, is used at each split point. The tree is then grown until it reaches a maximum depth or minimum node size.

A safeguard should be placed in the code so that each terminal node has both treatment levels. To estimate ITE from the RFIT model fit, the potential outcomes model is used: the estimated response is obtained for each individual for when they do or do not receive the treatment. To do this we subgroup observations in a terminal node into treatment and control groups and then take the average response of each subgroup. The difference of these averages is assigned to each individual in that terminal node as the estimated treatment effect. These estimates are averaged across the entire forest to give the estimated treatment effect for each individual.

2.2 Generalized Propensity Scores

In observational studies, the treatment is not randomly assigned, which is the cause of selection bias. If not properly accounted for, selection bias could lead to biased estimates of the treatment effect. Rosenbaum and Rubin (1983) show that, conditioning on propensity scores, treatment assignment and covariates are independent, which allows us to attain unbiased estimates of the treatment effect. This holds true under two assumptions: first, the potential outcomes are independent of the treatment assignment given the covariates, $\{Y(0), Y(1)\} \perp\!\!\!\perp Z|X$ with X being a vector of measured covariates; and second, each individual has a non-zero chance of receiving each treatment option. These two assumptions together are called strong ignorability.

When the treatment assignment is binary, the propensity score is defined as the conditional probability of being assigned treatment given the baseline covariates. In this paper we will denote the propensity score as e, defined as $e = P(Z = 1|X)$ where Z denotes the treatment variable and X denotes a vector of covariates. Propensity scores can be estimated using any classification machine learning method. Popular approaches include logistic regression and random forest.

The propensity score method can be extended to situations when the treatment variable has more than two levels (Imbens, 2000). We refer to the propensity score when there is more than two treatment levels as the generalized propensity score, defined as $r(z, X) = Pr(Z = z|X)$, where $z \in \{1, 2, \ldots, m\}$ denotes the value of a treatment variable with m levels. Generalized propensity scores can be estimated using any multiple classification machine learning methods such as multinomial logistic regression.

2.3 Causal Effect Random Forest of Interaction Trees for Multiple Treatments

RFIT makes the assumption that treatment is randomly assigned to each observation, so applying RFIT to observational study data could lead to biased estimates. To solve this problem RFIT was extended to causal effect random forest of interaction trees (CERFIT) to handle observational study data (Li et al., 2022). There are three changes made to RFIT to allow it to handle observational study data, all of which involve using propensity scores during the tree growing and prediction processes. The first change to RFIT is in the bootstrap sampling prior to growing each tree. In RFIT, each observation has an equal probability of being sampled, but in CERFIT a weighted bootstrap sample is used with weights given by $w_i = \frac{Z_i}{e_i} + \frac{1-Z_i}{1-e_i}$, that is inverse probability of treatment weighting (IPTW). Such weighting allows observations that were unlikely to receive a treatment option they ultimately did receive, to have a higher chance of being chosen in each bootstrap sample. This weighted bootstrap scheme helps adjust for the self-selection bias present in observational studies. The second change to RFIT is to include propensity scores in the linear model used to determine the best split. Modifying equation 2, we consider the model

$$Y_i = \beta_0 + \beta_1 I(Z_i = 1) + \beta_2 I(X_{ij} < c) + \beta_3 I(Z_i = 1) I(X_{ij} < c) + \beta_4 e_i + \varepsilon_i, \tag{3}$$

where e_i denotes the propensity score for the ith observation. The third change to RFIT is made when calculating the predicted values in each terminal node. Instead of using the node average for each treatment group to calculate the ATE in each terminal node, a weighted average is used where each observation is weighted according to IPTW. The weighted ATE is calculated as

$$ATE_w = \frac{\Sigma_i w_i I(Z_i = 1) Y_i}{\Sigma_i w_i I(Z_i = 1)} - \frac{\Sigma_i w_i I(Z_i = 0) Y_i}{\Sigma_i w_i I(Z_i = 0)}, \tag{4}$$

the difference in the weighted average response between treated and control subjects in the terminal node, with weights assigned according to IPTW.

Li et al. (2022) also extends RFIT to multiple treatments. Three changes were implemented to accommodate multiple treatments. The first change is to adopt the generalized propensity score method so that the weights applied for bootstrap sampling and for making predictions in terminal nodes are given by $w_i = r(z_i, X_i)^{-1}$. The second change is that at each potential split, two random treatment levels are selected. The treatment thus becomes a binary variable, and only subjects that were assigned one of the two randomly chosen treatments are used to choose the best split. Because there are typically a lot more splits in the random forest than the number of possible pairs of treatments, every pair of treatments should be selected, most likely multiple times, across the entire forest. The modified model for splitting is

$$Y_i = \beta_0 + \beta_1 I(Z_i = u) + \beta_2 I(X_{ij} < c) + \beta_3 I(Z_i = u) I(X_{ij} < c) + \beta_4 r(z_i, X_i) + \epsilon_i \tag{5}$$

for $i \in \{1, \ldots, n : Z_i = u \text{ or } v\}$, where u and v denote the two randomly selected treatments. The third, more minor, change is made when making predictions in terminal nodes. Instead of grouping the observations in each terminal node into treatment and control groups, they are grouped according to each level of treatment. As such, an observation will have a predicted response for each category of treatment, which can be used to obtain the estimated treatment effect comparing any category against the reference category. A drawback of this method is that only a subset of the data is used at each internal node to select the best split, which in turn may negatively affect the prediction performance of the random forest.

2.4 Proposed CERFIT for Ordered Treatment

In this paper we propose a modification of Li et al. (2022) CERFIT in order to handle ordered treatments. In many applications (such as in our motivating example of tutoring center visits), treatment can be represented as an ordered variable. Even if a treatment variable is continuous on the real line (which is rare in practice), percentiles of the treatment can be used to apply the proposed method. Our proposed algorithm takes advantage of the randomization scheme in RF by randomly dichotomizing treatment into a binary variable at each split. Since the random dichotomization is performed at every split of each tree, all treatment levels may be selected as the dichotomization point and dispersed over all the splits across the forest. The modified model for splitting is

$$Y_i = \beta_0 + \beta_1 I(Z_i < l) + \beta_2 I(X_{ij} < c) + \beta_3 I(Z_i < l) I(X_{ij} < c) + \beta_4 e_i + \epsilon_i, \tag{6}$$

where l denotes the dichotomization point and $e_i = \sum_{z<l} r(z, X_i)$ sums up generalized propensity scores according to how the treatment variable is dichotomized at the split. Following the `randomForest` R package, the default $mtry$ value is set at $p/3$ for the regression problem at hand, where p is the number of predictors. Equation 6 is similar to equation 5, but with a substantial improvement: all the data are used in equation 6, instead of only a subset of data corresponding to the two randomly selected levels of treatment values. A byproduct of this change is that the propensity score used is the one corresponding to how treatment is dichotomized. These changes recognize the ordered nature of the treatment variable and hence allow the proposed method to perform better than Li et al. (2022) when the treatment variable is an ordered one. Furthermore, using all of the data at each split point allows us to grow fewer trees than CERFIT for multiple treatment.

Binary Response: Equation 6 applies when the response variable is continuous. When the response variable is binary, a corresponding logistic regression model may be fit to search for the best split. In particular, the left hand side of equation 6 may be replaced by $Y_i' = \text{logit } Pr(Y_i = 1) = \log Pr(Y_i = 1)\{1 - Pr(Y_i = 1)\}^{-1}$. As before, the best split is the one that maximizes the squared test statistic for $H_0 : \beta_3 = 0$.

Numerical Stability: A criticism of RFIT is that the best split is chosen based on a single covariate without controlling for other covariates, as there is a possibility that the optimal split could be affected by other covariates (Alemayehu et al., 2017). To control for this, the left hand side of equation 6 may be replaced by residuals, denoted as $\tilde{Y}_i$, from a linear regression model where all covariates (but not treatment) are included. According to Li et al. (2022), using residuals from a regression model in place of the original responses increases the prediction accuracy of CERFIT. Similarly, when the response is a binary variable, a logistic regression model may be fit and the residuals may be used as a continuous response when growing the forest.

Weighting the bootstrap sample by IPTW may induce correlated data as observations with smaller values of generalized propensity scores are sampled more often than others. To account for this correlation, the robust standard error is used when calculating the test statistic for β_3.

For each terminal node, the predicted outcome under each treatment is given by the weighted average based on the subjects in the terminal node, with weights given by the inverse of generalized propensity scores. The estimated treatment effect for the terminal node is the difference in the predicted outcome between two treatment levels, and may be obtained for any pair of treatments. Any subject that ends up in the terminal node is assigned this estimated treatment effect as their ITE. The ITE for a subject from the forest is the average treatment ef-

Table 1: The proposed causal effect random forest of interaction trees (CERFIT) algorithm for ordered treatments with binary response.

1. Estimate generalized propensity score $r(z_i, X_i)$ *and calculate inverse probability weight* w_i.
2. Fit a logistic regression model with all covariates (but not treatment), and obtain residuals $\widetilde{Y}_i$.
3. Draw bootstrap samples from the data using w_i *as sampling weights.*
4. Grow an interaction tree based on each weighted bootstrap sample.
−4.1 At each node, randomly select $mtry$ *of the total* p *covariates from which to determine the splitting rule. The default value of* $mtry$ *is set at* $p/3$.
−4.2 Randomly dichotomize the treatment variable.
−4.3 Among the $mtry$ *covariates selected, the optimal split is identified by maximizing the robust squared Wald test statistic for testing* $H_0 : \beta_3 = 0$, *in equation (6) using residuals* $\widetilde{Y}_i$ *as outcome.*
−4.4 Repeat steps 4.1 *to* 4.3 *until reaching a pre-specified stopping rule (e.g., maximum tree depth, minimum terminal node size).*
5. Repeat steps 3 and 4 for $ntree$ *trees as desired, with a default value set at* 500.

fect across all trees. The optimal treatment for each subject is the one associated with the largest treatment benefit. As in the RF algorithm (Breiman, 2001), predictions of the ITE should be based on an out-of-bag (OOB) sample, which is composed of the observations left out from the weighted bootstrap sample. The algorithm for the proposed method is presented in Table 1.

3 Simulation Studies

To evaluate the effectiveness of the proposed method, simulation studies were conducted. Results will be presented in terms of variable importance rankings, classification accuracy (%) for the optional ITR, and prediction accuracy for the outcome measured by mean squared error (MSE).

The structure of the simulation studies was modified from the simulation studies found in Zhu et al. (2015) and Li et al. (2022). Three different simulation scenarios were used and for each scenario a sample size of $n = 1000$ was used for training and testing. All simulations were done using the R programming language (R Core Team, 2021). The simulated data included ten covariates with each being independently generated from a standard normal distribution.

The treatment variable was simulated from a binomial distribution with $n = 5$, $Z \sim Binom(5, m)$, so that each subject could receive a treatment option from $\{0, 1, \ldots, 5\}$. The probability of success (treatment), denoted as m, depends on the covariates and varies from scenario to scenario. The complexity of m increases from scenario 1 to scenario 3, see equations 7, 8, and 9 below.

The response, Y, is modeled with the equation $y(X, Z) = h(X) + g(X, Z) + \varepsilon$, where $h(X)$ depends on X only and hence does not affect the ITR, and $g(X, Z)$ contains interactions between X and Z and hence defines the optimal ITR. For all three scenarios $h(X)$ remains the same, and is given by $h(X) = .3X_1 + .36X_2 + .73X_3 - .2X_4 + .1X_5 + .25X_8^2$. On the other hand, the function g varies in each scenario. Scenario 1 features a simple non-linear treatment interaction effect, scenario 2 features a tree-like treatment by interaction effect, and scenario 3 features a complex non-linear treatment by interaction effect. The functional forms of both m and g for the three scenarios are as follows.

Scenario 1

$$logit(m) = .3X_1 + .65X_2 - .35X_3 - .4X_4 \tag{7}$$

$$g(X, Z) = (X_1 + X_2) * log(Z + 1)$$

Scenario 2

$$logit(m) = .3I(X_1 > .5) + .65I(X_2 < 0) - .35I(X_3 > .75) - .4I(X_4 < .75) \tag{8}$$

$$g(X, Z) = I(Z \geqslant 3) * (X_1 * X_2) + (Z \leqslant 4) * (X_1 + X_2) + I(2 \leqslant Z \leqslant 4) * I(X_1 * X_2 < 0)$$

Scenario 3

$$\begin{aligned} logit(m) = {} & .3X_1 + .65X_2 - .35X_3 - .4X_4 + .65I(X_1 > 0)I(X_1 > 1) \\ & + .3I(X_1 > 0)I(X_4 > 1) - .65I(X_2 > 3)I(X_3 > 0) \end{aligned} \tag{9}$$

$$g(X, Z) = -.05(Z - \frac{X_1 + X_2}{4} - 3)^2 + .5 * Z * I((X_1 - X_2) > 0)$$

The proposed method was compared to Bayesian additive regression trees (BART) (Chipman, 2010) and CERFIT for multiple treatments (Li et al., 2022). BART was implemented using the BART R package (Sparapani et al., 2021). The two CERFIT models, for multiple treatments and ordered treatments respectively, were implemented using the CERFIT R package (Thorp et al., 2022).

In this paper, we use minimal depth (Ishwaran et al., 2012) to quantify variable importance for its computational efficiency. Minimal depth for any variable in a tree is defined as the smallest distance between a node, where this variable is used as the splitting variable, and the root node. If the variable is used more than once in the tree, the minimal depth is the depth between the node that splits on this variable and is also the closest to the root node. If the variable is not used in the tree, the minimal depth is given as the max depth of the tree, plus one. The average minimal depth across the entire forest is used to measure variable importance. Variables with lower average minimal depths are more important. Figure 1 presents box plots of average minimal depths. Each box plot is based on the average minimal depths from 100 simulated data sets with 500 trees grown for each data set. In each of the three scenarios the proposed CERFIT method for ordered treatments is able to correctly identify the most important variables that impact treatment effect despite the differing levels of complexity in each scenario. Note that in all three scenarios, the only important predictors that impact treatment effect are X_1 and X_2.

Figure 2 shows the correct classification rate (%) for the optimal treatment. Comparisons are made to BART and CERFIT for multiple treatment. Higher rates of correct classification for the ITR indicate better performance. For all three scenarios, and on average, both of the CERFIT methods perform better than BART. These results show that in order to recommend the optimal treatment, it is crucial to incorporate the propensity score into the model when using observational study data. In comparison to Li et al. (2022), the proposed method is a winner in all three scenarios, indicating that preserving the ordered nature of the treatment and using all available data at each split does lead to better predictions when the treatment variable is truly ordered.

Figure 3 presents prediction accuracy for the outcome as measured by the mean squared error (MSE). Again the proposed method is compared to BART and CERFIT for multiple

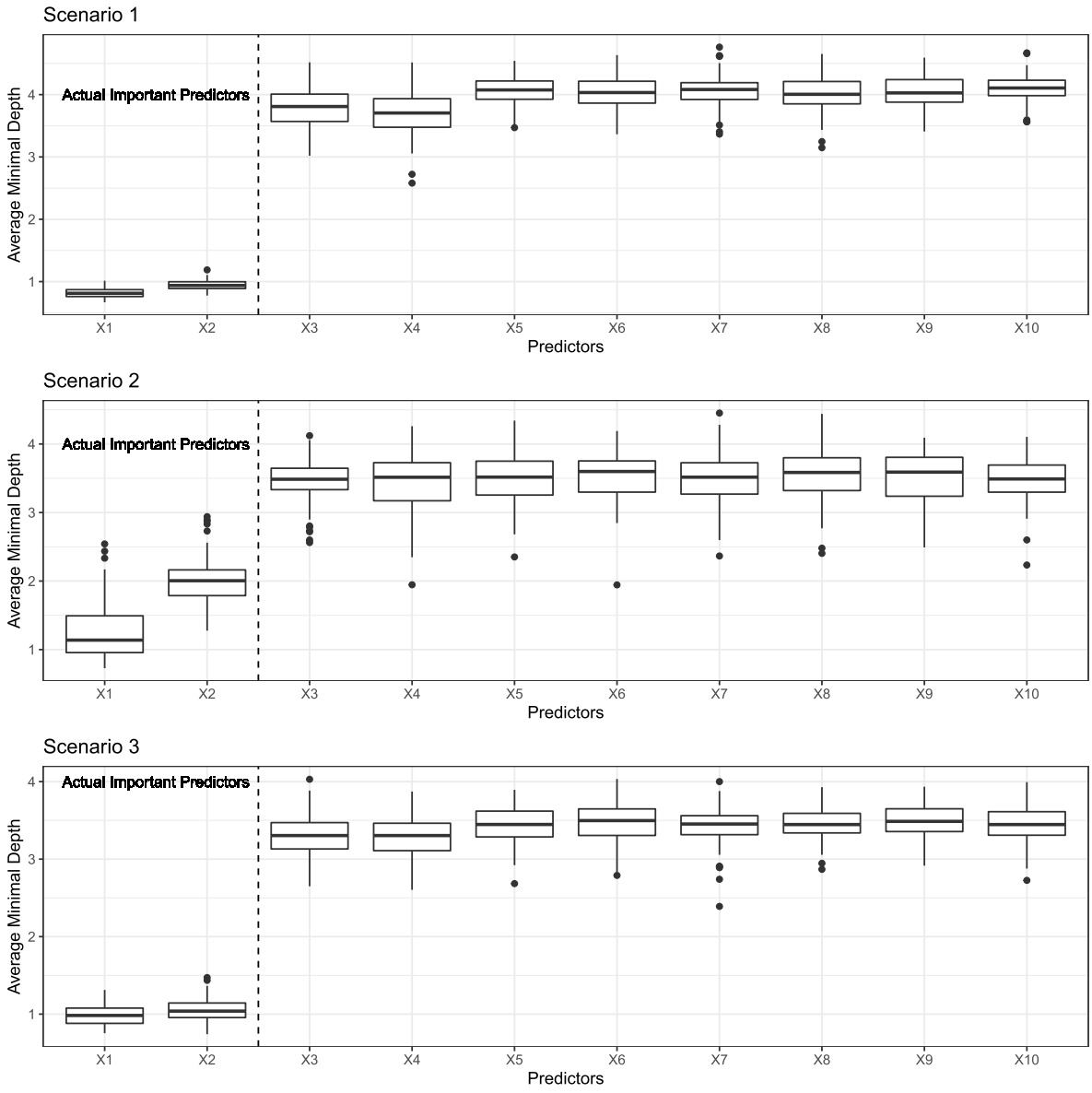

Figure 1: Variable importance plots for three different scenarios based on 100 random forests, with 500 trees in each forest. Smaller average minimal depths represent more important variables. Variables X_1 and X_2 are the only important predictors that impact treatment effect in each scenario.

treatment. Here smaller values of MSE indicate better accuracy. For all three scenarios, the proposed method has better accuracy, on average, than CERFIT for multiple treatments (Li et al., 2022). In comparison to BART, the proposed method has better accuracy than BART in two of the three scenarios (scenarios 1 and 2). In scenario 3, BART has slightly lower median MSE than the proposed method.

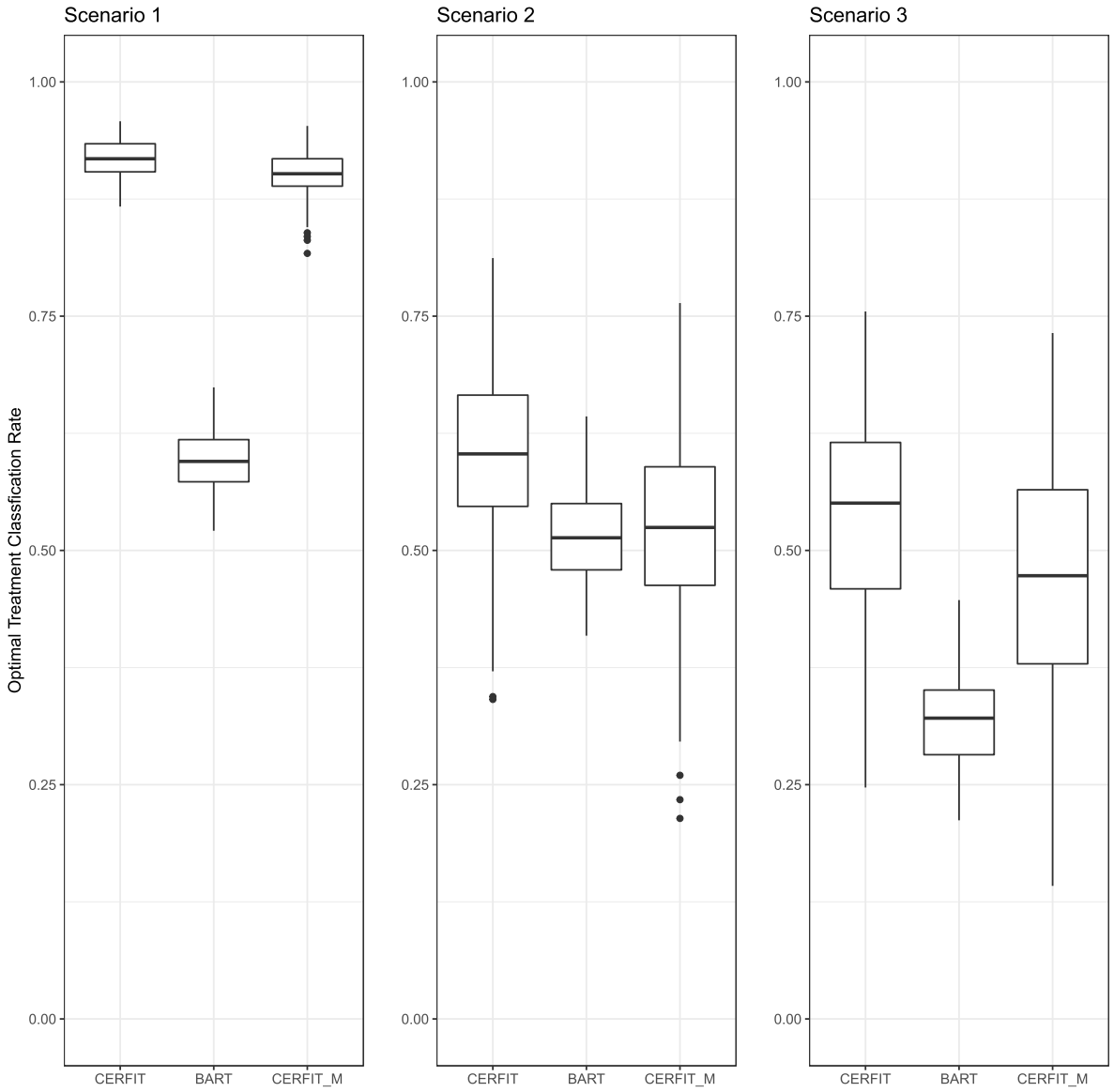

Figure 2: Box plots of correct rate (%) of optimal treatment classification for three different scenarios based on 100 random forests, with 500 trees in each forest. Higher percentages indicate better performance. Three methods are compared: the proposed causal effect random forest of interaction trees (CERFIT) method, Bayesian additive regression trees or BART (Chipman, 2010), and CERFIT for multiple treatments (CERFIT_M) (Li et al., 2022).

4 Application to Educational Data

This section presents an application of the proposed method to student success data, collected from students enrolled in an introductory statistics course at San Diego State University. The data, from 1386 students, include 15 covariates such as age, high school GPA, and SAT composite score; see Table 2 for the full list of variables along with descriptive statistics for each. Underrespresented minority (URM) is defined, by the California State University system, as

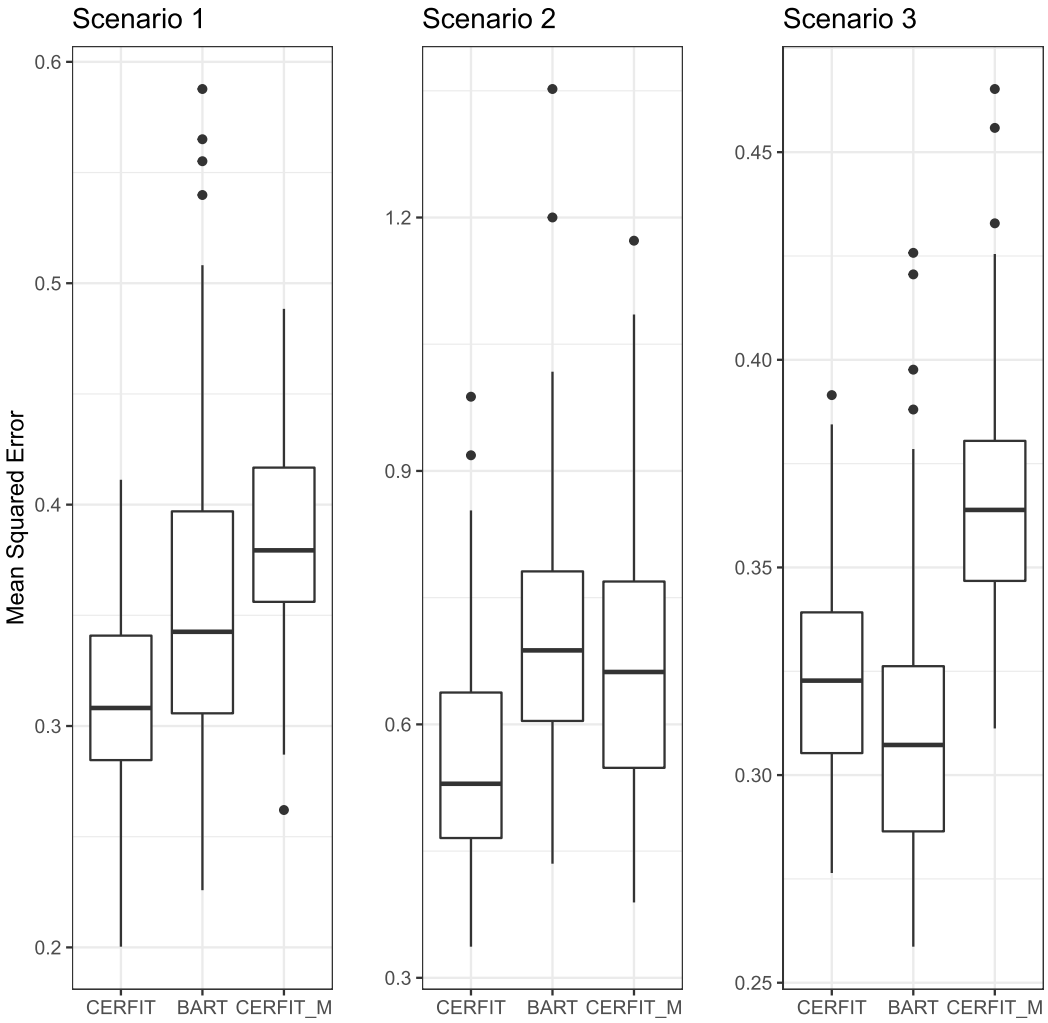

Figure 3: Box plots of mean squared error (MSE) of predicted outcome for three different scenarios based on 100 random forests, with 500 trees in each forest. Lower values of MSE indicate better performance. Three methods are compared: the proposed causal effect random forest of interaction trees (CERFIT) method, Bayesian additive regression trees or BART (Chipman, 2010), and CERFIT for multiple treatments (CERFIT_M) (Li et al., 2022).

people who identify as African American, Latinx, or Native American. The Compact Scholar Program is offered to students from underprivileged communities (the word "compact" refers to the agreement between SDSU and a local school district). The outcome variable is whether the student passes an introductory statistics course, defined as earning a grade of C or better. The treatment being evaluated is the number of visits to the Math and Stat Learning Center (MSLC) on campus. The MSLC is a free tutoring center for students who wish to get help in mathematics or statistics. We are interested in learning the effect of the number of visits for the class on the probability of passing the course. Because students voluntarily choose to visit the MSLC or not, these data are observational and the covariates that are associated with getting a higher grade might also be associated with a student choosing to go to the MSLC. In order to

Table 2: The variables included in the educational data set. The summary statistics given in the last column are sample size (proportion) for qualitative variables and mean (standard deviation) for quantitative variables.

	Variable Name	Value	Summary
Categorical Predictors	Semester	S18	702 (0.51)
		S19	684 (0.49)
	Gender	Female	837 (0.60)
		Male	549 (0.40)
	Underrepresented Minority	No	833 (0.60)
		Yes	553 (0.40)
	Student Level	Freshman	165 (0.12)
		Sophmore	998 (0.72)
		Junior	181 (0.13)
		Senior	42 (0.03)
	Education Opportunity Program	No	1308 (0.94)
		Yes	78 (0.06)
	Compact Scholar Program	No	1233 (0.89)
		Yes	153 (0.11)
	Scholarship Program	No	1319 (0.95)
		Yes	67 (0.05)
	On Campus Housing	No	900 (0.65)
		Yes	486 (0.35)
	STEM	No	950 (0.69)
		Yes	436 (0.31)
	Campus GPA	A	437 (0.32)
		B	675 (0.49)
		C	222 (0.16)
		D	51 (0.04)
		F	1 (0.0)
Continuous Predictors	High School GPA		3.69 (0.30)
	SAT Composite		1169.46 (130.79)
	Total Units Enrolled		15.17 (2.04)
	Age		18.62 (0.88)
	Grade		778.20 (171.58)
Outcome Variable	Pass	No	271 (0.20)
		Yes	1115 (0.80)

obtain unbiased estimates of the treatment effect, the propensity score method needs to be used for causal inference. The main goal of the analysis is to identify the optimal number of visits for each student in an effort to maximize their chances of passing the class.

Table 3 presents the observed frequencies for the number of visits to the MSLC for the

Table 3: The observed frequencies for the number of visits to the Math and Stat Learning Center for the introductory statistics course.

# of visits	0	1	2	3	4	5	6	7	8	9	10
Frequency	1010	214	73	25	16	8	7	2	5	4	2
# of visits	11	12	13	14	17	18	20	22	23	25	
Frequency	5	3	2	2	2	1	1	2	1	1	

Table 4: Crude passing rate by number of visits to the Math and Stat Learning Center.

#Visits	#Students	Passing Rate
0	1010	0.78
1	214	0.90
2	73	0.85
3	25	0.88
4	16	0.88
5-7	17	0.88
8-11	16	0.88
12+	15	0.87

introductory statistics course. Note that the frequencies are rather small for any number of visits larger than 4. Consequently, the levels of the treatment variable are grouped into 0, 1, 2, 3, 4, 5-7, 8-11, and 12+ visits, so that each level has a sufficient sample size, see Table 4. Also included in Table 4 are the crude course passing rates (without adjusting for covariates) for each visit bracket. It can be seen from Table 4 that there does seem to be a difference in the passing rate between those students who did not go to the MSLC at all and those students who went at least once. But, there does not seem to be much difference among students who visited the MSLC at least once. However, the true benefits of MSLC attendance might depend on individual student characteristics. To this end, the proposed CERFIT method was applied to these data using 2000 interaction trees with an $mtry$ value of 5.

Figure 4 presents the predicted ITE based on OOB samples, versus number of visits. The ITEs are given relative to no visit to the MSLC. Almost all students would see a positive benefit from going to the MSLC once. Most students would also benefit from going to the MSLC up to 7 visits. However, only some students would benefit from going to the MSLC for 8 or more times (last two violin plots). The group with three visits is clearly the group having the smallest number of students who would receive a negative effect, and seems to have the highest chance of being a student's optimal treatment. Overall, the plots suggest that most students would benefit the most from up to 4 visits to the MSLC. A possible reason for this finding is that a few visits right before a midterm or final exam could really benefit a student. Another possibility could be that some students with prior poor performance or low aptitude for statistics may be attending more tutoring sessions but receiving less benefit. In terms of ITE variation, the group with one visit seems to have the smallest variance out of all visit groups, which may be due to the relatively large sample size (number of students in the group). For the same reason, the ITE for last four visit brackets have the largest variance since these groups have the smallest sample

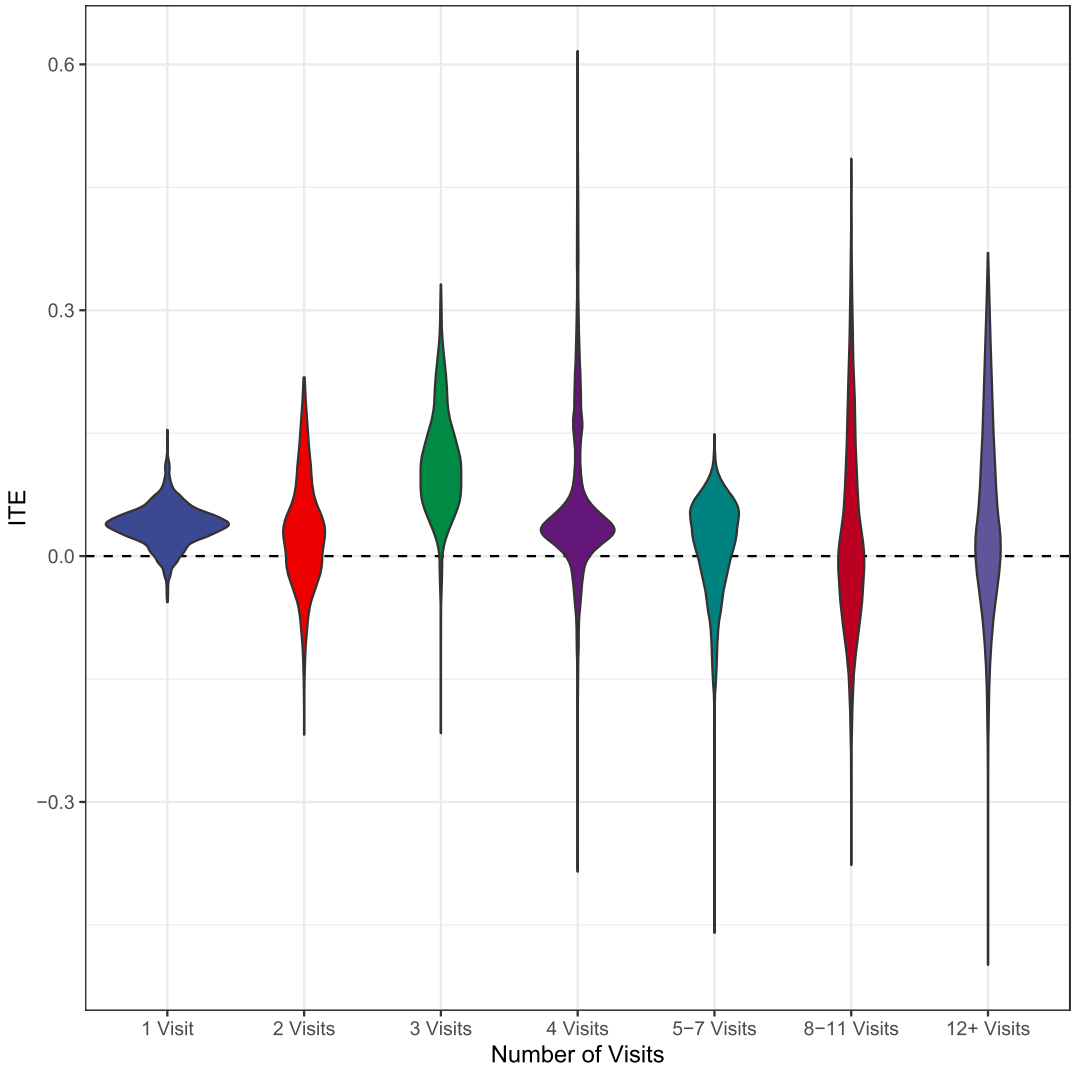

Figure 4: Violin plots of individualized treatment effect (ITE) for each student at each treatment level. The ITE's are given in terms of the probability of passing the class, compared to no visit to the Math and Stat Learning Center.

sizes.

Figure 5 presents average minimal depth for all the covariates used in the analysis, with smaller minimal depths indicating more important variables that impact the treatment effect. The results show that high school GPA, SAT composite score, and total number of enrolled units during the semester are the variables that impact the treatment effect the most. While SAT composite score and high school GPA are the usual suspects, the total number of enrolled units is not. Figure 6 shows the relationship between the predicted ITE and the number of visits to the MSLC, but broken down into three groups based on the number of units enrolled. The three groups correspond roughly to students that were less than full time (3-13 units; 259 students), full time (14-16 units or about 5 courses with 3 units for each course; 867 students), and more than full time (17 or more units; 260 students). The figure shows that, overall, students

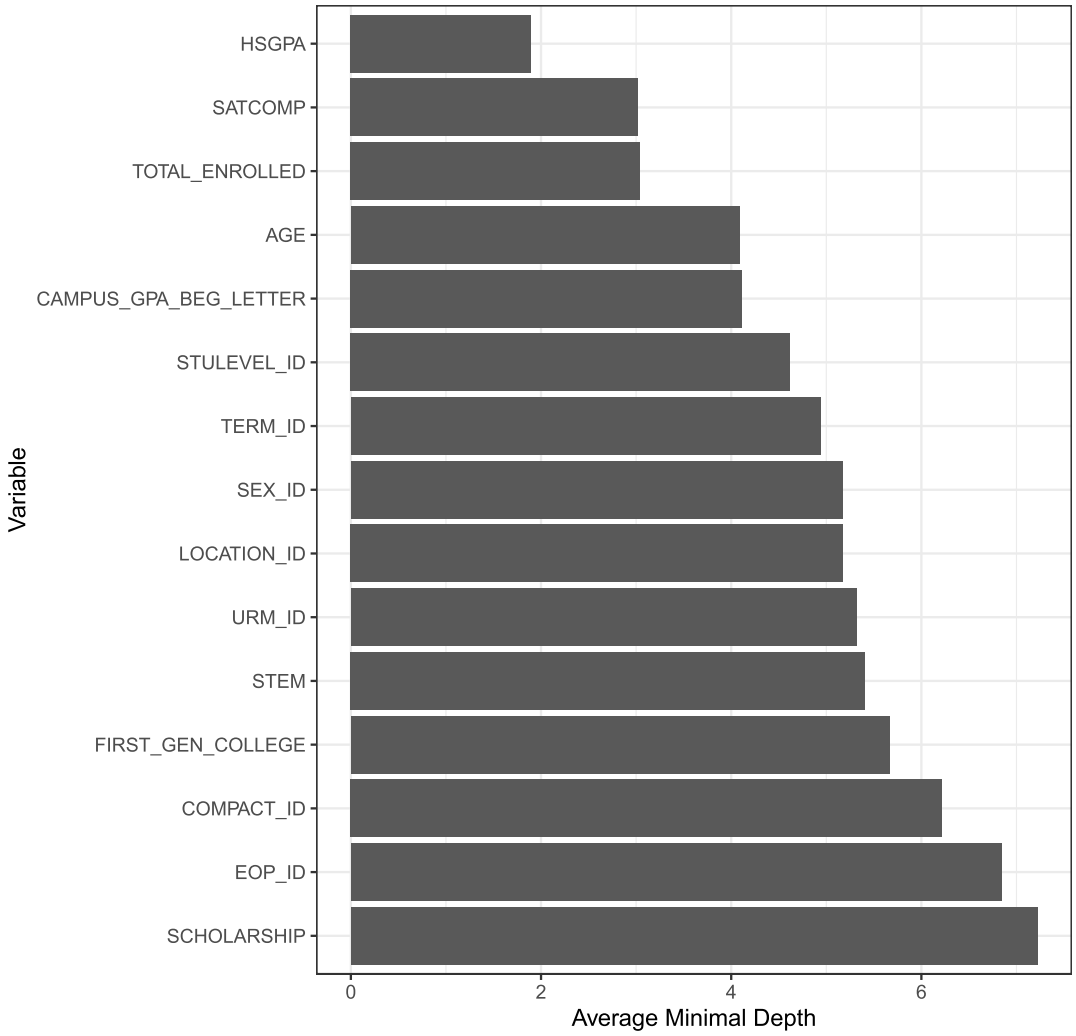

Figure 5: The average minimal depth of all covariates used in the analysis. Smaller average minimal depths signify more important covariates that impact the treatment effect.

enrolled in the least units (blue solid curve) received higher benefits from visits to the MSLC. While the curves for students enrolled in 3-13 and 14-16 units (blue solid and red dashed curves respectively) do not cross, the curve for students enrolled in 17 or more units (green dashed curve) starts out at the lowest position (smallest benefits from visits to MSLC) but then moves higher crossing both of the other two curves. This finding suggests that there is a group of "super-charged" students, who enrolled in a lot of courses, visited the MSLC many times, and still received great benefits from these visits. One can also see from this Figure that 3 visits to the MSLC seem to be optimal for most students.

Table 5 shows the average values of student characteristics and outcomes, grouped by their recommended optimal treatment regimes (i.e., optimal number of visits to the MSLC). The table shows that the majority of students were recommended between one to three visits to the MSLC, and yet a sizable number of students were recommended four or more visits. No student was recommended zero visits indicating that every student could benefit from visiting the MSLC

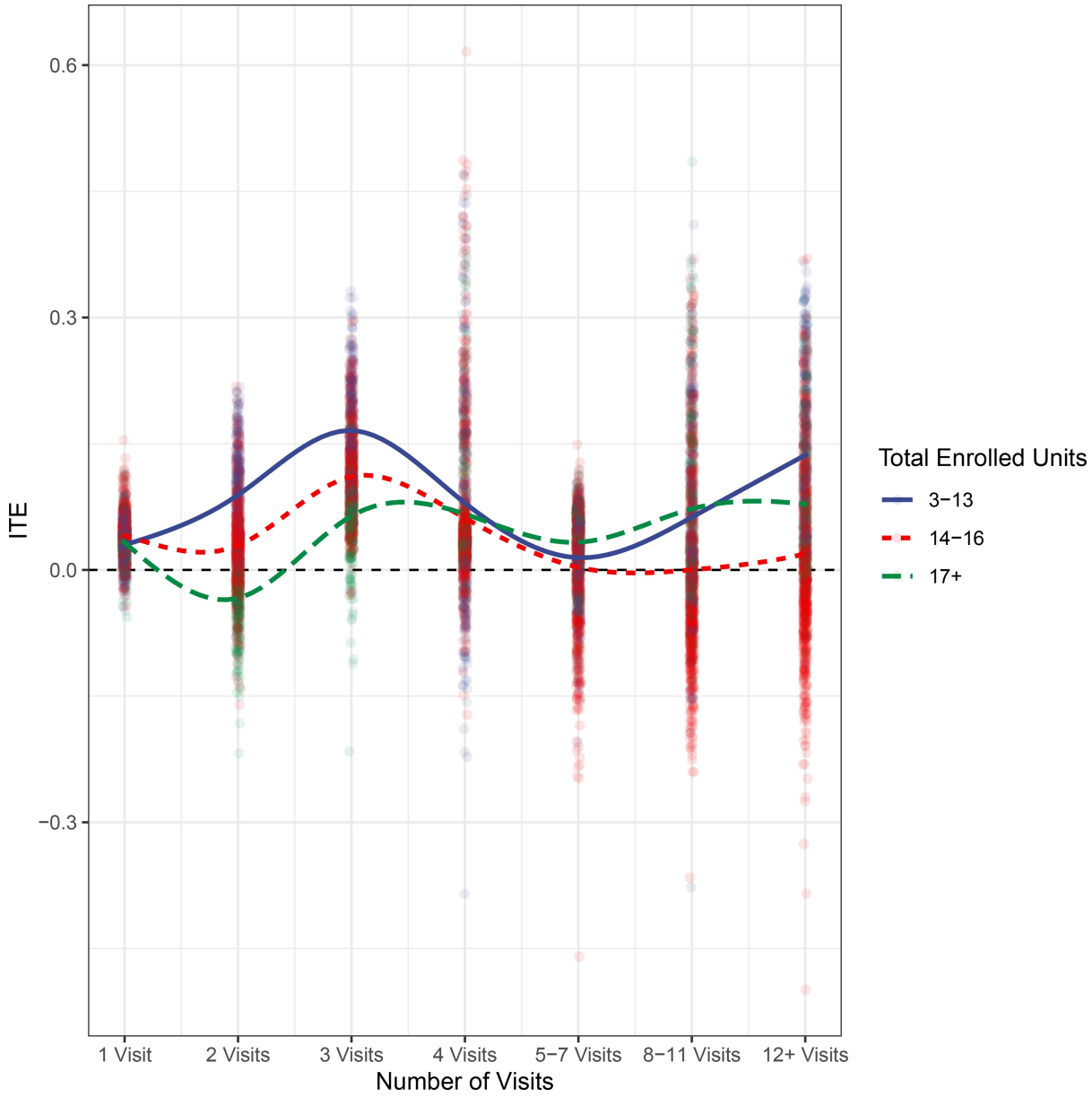

Figure 6: Relationship between the individualized treatment effect (ITE) and the number of visits to the Math and Stat Learning Center, broken down by three groups of students with differing number of total enrolled units during the semester.

at least once. The table also shows that students who were recommended more visits had, on average, a lower high school GPA and lower campus GPA. In addition, more URM students were recommended more visits on average. From the outcome side, students who were recommended the fewest (1-3) visits had a higher average course grade (out of 1030), and a higher probability of passing the course, compared to the students recommended 4 or more visits. Interestingly, even though students who were recommended more visits had lower grades, they did receive higher benefits (ITE) from visits to the MSLC.

Table 5: Average values of student characteristics and outcomes, grouped by their recommended optimal treatment regimes (i.e., optimal number of visits to the MSLC).

	1-3 Visits	4-7 Visits	8+ Visits
Number of Students	912.00	218.00	256.00
HSGPA	3.80	3.61	3.39
SATCOMP	1170.79	1179.59	1156.09
Total Enrolled	14.98	15.32	15.76
Campus GPA	3.24	2.72	2.82
URM	0.38	0.42	0.45
First Gen	0.13	0.15	0.12
Grade	803.51	715.76	741.24
Pass	0.86	0.64	0.73
ITE	0.12	0.19	0.18

One of the reasons to study ITE and ITR is that the same treatments may have very different effects on different subjects. Figure 7 shows plots of ITE for ten randomly selected students. We can see that students did experience drastically different responses to the same levels of treatment. Circles in the plots indicate the observed (or actual) number of visits to the MSLC. If a student's plot does not have a circle around any of the points, it means that the student did not visit the MSLC (the actual number of visits was zero). For each plot, the number of visits with the highest ITE value is the optimal treatment regime for the student. Observed values of selected variables for these randomly selected students are given in Table 6. Students one, six, and ten see very similar ITE curves, where we see an overall increasing trend as the number of visits increases. This finding can be explained in part by the students' similar values in both SAT composite score and high school GPA. What is interesting is that these students had very different final grades, explained at least partially by their chosen treatment level. While students one and ten did not visit the MSLC at all, student six went eight to eleven times, which happened to be the recommended optimal number of visits for this student. Note that students with low high school GPA and low SAT composite score really benefit from a large number of visits to the MSLC. Students three, four, and nine see very similar trends in that as their number of visits increases beyond 3, their ITE seems to have an overall decreasing trend. As these students all have very good high school GPA and a full-time course load, their time might be spent more effectively studying on their own beyond visiting the MSLC a limited number of times. Note that there is substantial variability in the ITE curves in Figure 7 due to limited sample size for the study, especially for the number of students who visited the MSLC for three or more times. When using these curves for student advising, it is important to consider the general trend in ITE as well as other factors such as student time constraints.

5 Discussion

In this paper we propose a novel method based on RFIT in order to estimate ITE and provide ITR for observational study data with multiple ordered treatment levels. In order to attain unbiased estimates of the ITE, generalized propensity scores are incorporated in the tree building process as well as in the prediction procedure based on the terminal nodes. Note that in many applications, treatment can be represented as an ordered variable, and even when the treatment

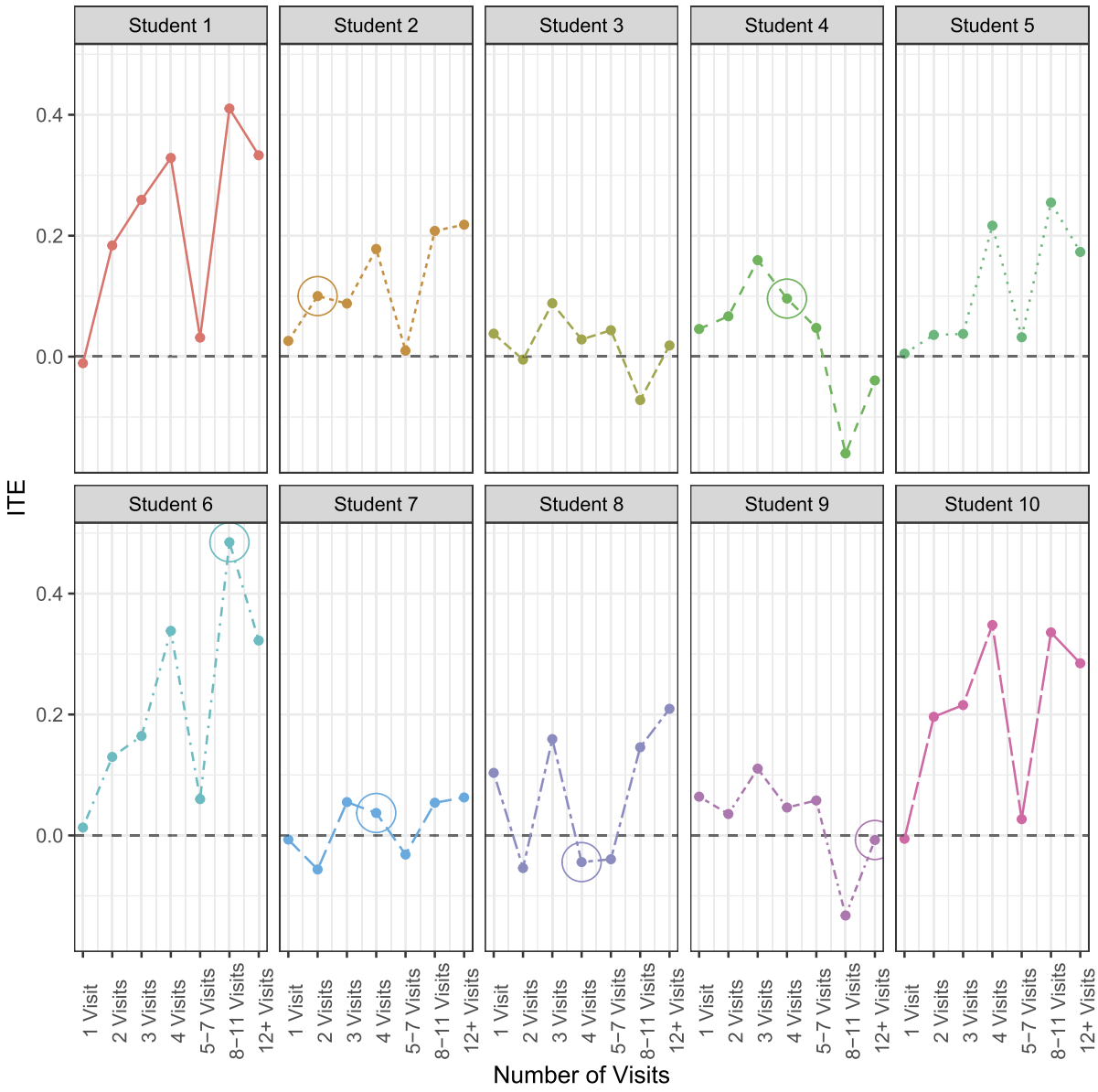

Figure 7: Estimated individualized treatment effect (ITE) versus number of visits to the Math and Stat Learning Center for 10 randomly selected students.

is continuous one may use its percentiles as ordered levels. A R package (Thorp et al., 2022) that implements RFIT for binary, multiple ordered, and multiple unordered (categorical) treatments, applicable to both RCT and observational data, has also been developed and made publicly available. We are currently working on an extension of the RFIT algorithm to a time-to-event response.

The results from our simulation studies show that the proposed method is effective and has better comparative accuracy compared to two existing methods: BART (Chipman, 2010) and the original CERFIT (Li et al., 2022). These two methods were chosen because BART was the most competitive method among various methods that Li et al. (2022) evaluated in a comprehensive

Table 6: Values of outcomes and selected covariates for the 10 randomly selected students in Figure 7.

Student	SAT	HSGPA	#Units	#Visits	URM	1st Gen	Grade	Pass
1	930	3.19	6	0	1	0	453.51	0
2	1070	3.23	12	2	1	0	1017.81	1
3	1340	3.94	16	0	0	0	900.13	1
4	1010	4.03	15	4	0	1	784.65	1
5	1260	3.09	18	0	1	0	250.49	0
6	970	3.28	17	8-11	1	0	810.76	1
7	1350	3.46	16	4	0	0	957.93	1
8	850	3.41	17	4	1	1	478.83	0
9	970	3.96	16	12+	1	0	792.06	1
10	940	3.21	12	0	0	0	619.95	0

simulation study, while the CERFIT method of Li et al. (2022) is the method we modified in order to handle ordered treatments more effectively. Using minimal depth as the variable importance measure, the proposed method was able to correctly identify the most important variables that impact the treatment effect. Note that variable importance in CERFIT measures the ability of variables to impact the treatment effect. Such a variable importance measure is not easily attained from BART. The variable importance results from Li et al. (2022) are very similar to those from the proposed method for ordinal treatments.

In terms of prediction accuracy, the simulation results show that the proposed method was able to attain a higher classification accuracy for the optimal treatment than BART (Chipman, 2010) and CERFIT for multiple treatments (Li et al., 2022) in all scenarios considered. In addition, the proposed method attained better prediction accuracy, signified by lower MSE for the outcome prediction, than BART and CERFIT for multiple treatments. This finding held in all scenarios except for one, where BART had a slightly lower median MSE. Note that the proposed method is designed specifically for predicting ITE, while BART is optimized for outcome prediction. The fact that BART does not account for the observational nature of the data, beyond adjusting for covariates in the tree growing process, probably hurt its performance when applied to observational study data.

The proposed method was applied to student success data from an introductory statistics course at San Diego State University. The results indicated that high school GPA, SAT composite score, and total units enrolled during the semester were the most important variables that impacted the effect of treatment (visits to the MSLC) on the probability of passing the course. The results also showed that every student would benefit from going to the MSLC at least once, and that for a majority of students the optimal number of visits to the MSLC was three. The reason for this last finding might be that visiting the MSLC right before a midterm or final exam could be very effective for students. Because of the small number of students going to the MSLC for more than 4 times, the higher visit numbers were collapsed into a few brackets. Note that even with grouping, the sample size in each bracket (with the smallest one at 15 students) is still not large. A larger data set could allow us to estimate the ITE for each distinct number of visits and reduce the variance of the estimates.

Work is currently being carried out that takes into account when visits to the MSLC take

place, relative to the semester and assessment timelines, as the timing of the visits might also impact their benefit. In addition, a simulation study is also currently underway to investigate the comparative performance of the proposed and competing methods when the strong ignorability assumption is violated, since real data may often have unmeasured confounders.

Funding

This research was supported in part by the National Science Foundation grant 1633130.

References

Alemayehu D, Chen Y, Markatou M (2017). A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical Methods in Medical Research*, 27(12): 3658–3678.

Breiman L (2001). Random forests. *Machine Learning*, 45: 5–32.

Breiman L, Friedman J, Stone C, Olshen R (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Chipman H (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4.

Dusseldorp E, Mechelen I (2013). Qualitative interaction trees: A tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*, 33: 219–237.

Imbens G (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87: 706–710.

Ishwaran H, Kogalur U, Gorodeski E, Minn A, Lauer M (2012). High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105: 205–217.

Li L, Levine R, Fan J (2022). Causal effect random forest of interaction trees for learning individualized treatment regimes with multiple treatments in observational studies. *Stat*, 11(1), e457.

Lipkovich I, Dmitrienko A, Denne J, Enas G (2011). Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30(21): 2601–2621.

Rosenbaum P, Rubin D (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41–55.

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688–701.

Sparapani R, Spanbauer C, McCulloch R (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, 97(1): 1–66.

Su X, Peña A, Liu L, Levine R (2018). Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Statistics in Medicine*, 37(17): 2547–2560.

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

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Thorp J, Li L, Fan J (2022). CERFIT: Causal Effect Random Forest of Interaction Trees. https://cran.r-project.org/web/packages/CERFIT/index.html.

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.

Zhu Y, Coffman DL, Ghosh D (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, 3(1): 25–40.