EFFICIENTLY EXPLORING MULTILEVEL DATA WITH RECURSIVE PARTITIONING

Daniel Patrick Martin
Enfield, CT

M.A. Quantitative Psychology, University of Virginia, 2013
B.S. Applied Mathematics, University of Rhode Island, 2011
B.A. Psychology, University of Rhode Island, 2011

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Psychology
University of Virginia
August, 2015

Timo von Oertzen
(co-chair)

Sara E. Rimm-Kaufman
(co-chair)

Brian A. Nosek

Vivian C. Wong
(Curry School of Education)

ABSTRACT

There is an increasing number of datasets with many participants, many variables, or both, found in education and other areas that commonly have complex, multilevel data structures. Once initial confirmatory hypotheses are exhausted, it can be difficult to determine how best to explore these datasets to discover hidden relationships that could help to inform future research. Typically, exploratory data analysis in these areas are performed with the very same statistical tools as confirmatory data analysis, leading to potential methodological issues such as increased rates of false positive findings. In this dissertation, I argue that the utilization of a data mining framework known as recursive partitioning offers a more efficient means to perform exploratory data analysis to identify variables that may have been overlooked initially in the confirmatory data analysis phase. By adopting such a non-parametric approach, researchers can easily identify the extent to which all variables are related to an outcome, rather than rely on null hypothesis significance tests as a strict dichotomization of whether a given variable is "important" or "unimportant." This dissertation evaluates the feasibility of using these methods in multilevel contexts commonly found in social science research by using both Monte Carlo simulations and three applied datasets. Based on these results, a set of best practices was constructed and disseminated via a small workshop given to applied researchers. Feedback from these researchers helped lead to a publicly available tutorial and R package to assist others interested in adding this technique to their own statistical toolbox.

We're here to help each other get through this thing, whatever it is.

— Mark Vonnegut

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my primary advisor, Timo von Oertzen, for being so supportive throughout my graduate career. You always found interest in whatever side-projects I was working on, regardless of whether they had a quantitative focus or not. This encouragement towards interdisciplinary collaboration had a huge impact on my research and career pursuits, and for that I will forever be grateful.

Also, thanks to Sara Rimm-Kaufman and Brian Nosek for always taking time out of your busy schedules to teach me a thing or two. If I can replicate even a fraction of the passion and dedication you both have, I know I will be in for a very successful career. I would also like to thank Bethany Teachman for trusting in my abilities, and both Vivian Wong and Erik Ruzek for providing so much support throughout my dissertation project.

I would like to give a special thanks to my friends in psychology and the Curry school, and to those in Durham, who gave me a home away from home. You are easily the smartest and kindest people that I know, and I consider myself lucky to have had the pleasure of getting to know you.

Finally, a special thanks should also be given to my family, for their unwavering love and support, and to Lauren, for keeping me well-fed, sane, and most importantly, happy. I certainly would not be anywhere close to where I am today if it was not for you.

Now, on to the next adventure.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Once upon a time statisticians only explored. Then they learned to confirm exactly - to confirm a few things exactly, each under very specific circumstances. As they emphasized exact confirmation, their techniques inevitably became less flexible. The connection of the most used techniques with past insights was weakened. Anything to which a confirmatory procedure was not explicitly attached was decried as 'mere descriptive statistics', no matter how much we had learned from it.

— Tukey (1977)

Consider the following scenario: An educational researcher just finished collecting data for a large grant examining how teacher-student interactions are related to student outcomes. When originally writing up the grant, the researcher had some specific hypotheses grounded in theory that the grant was designed to test. The dataset was cleaned and prepped, and the hypotheses were empirically tested. These tests yielded results providing support for some of the original hypotheses, while others were a little less clear. Naturally, the researcher had collected additional variables that may or may not be relevant to the outcome (e.g., teacher and student demographics), and is now interested in performing exploratory data analysis to examine these variables further. However, unlike the original hypotheses, theory does not have strong predictions for these additional variables. Interested in performing a more exploratory study with these data, the researcher now has to make decisions regarding which variables to include, how these variables might be related (linearly, quadratically, etc.), and if potential interactions might be necessary in these exploratory models. How does the researcher proceed?

This situation is becoming more common as there is an increasing number of datasets with a large number of participants, variables, or both, in education and other fields that often deal with multilevel data structures. With this increase of available information, it becomes necessary to be able to efficiently search a large parameter space to identify variables that might have been overlooked to uncover insights and help inform future research. Currently, this practice is typically accomplished in the social sciences "by hand." That is, the researcher in question will run multiple tests from a hypothesis testing framework with different model specifications and combinations of variables in order to determine which are the most important (Strobl, Malley, & Tutz, 2009).

While some argue that exploratory approaches do not need any type of correction as the results are preliminary (Schochet, 2008), performing exploratory data analysis solely with a hypothesis testing framework is still rife with statistical issues regarding the generalizability of such findings. For example, it is easy to blur the lines between what is confirmatory and what is exploratory when confronted with a large dataset with many possible quantitative decisions (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). In such situations, there is an increased chance of detecting spurious findings due to the large number of potential researcher degrees of freedom (Gelman & Loken, 2014; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). Even seemingly simple choices of what covariates to include or which distribution to specify for the outcome is enough to make researchers come to different conclusions regarding the statistical significance of a given variable when confronted with the same research question (Silberzahn et al., in prep).

To address this concern, more emphasis is being placed on replication (Open Science Collaboration, 2014, in press). Although many participating in this movement stem from social psychology in particular, this movement has also been echoed in other areas where complex multilevel data are commonplace, such as developmental psychology (Duncan, Engel, Claessens, & Dowsett, 2012) and education (Makel & Plucker, 2014). While replicability is a hallmark of good scientific practice (Nosek, Spies, & Motyl, 2012), it is more feasible to implement as a means to control for potential spurious effects caused by a large number of researcher degrees of freedom in some domains more than others (Finkel, Eastwick, & Reis, 2015). In many social psychology lab studies, for example, the preliminary nature of exploratory data analysis can simply be confirmed with the running of additional studies at relatively minimal cost. Studies that are run in settings as complex as a school system unfortunately do not have this luxury.

What is needed, then, is a more efficient means of exploratory data analysis that can help uncover insight while simultaneously controlling for spurious findings, have the ability to be implemented when theory might not dictate how the model should be specified, handle situations where complex, multilevel data structures are commonplace, and, arguably most important, be something that can be adopted by the average, applied researcher. Given that previous research has focused solely on performing confirmatory and exploratory research from a null hypothesis significance testing perspective (Berk, 2008; Finkel et al., 2015), identifying a potential solution to this problem seems almost infeasible. However, solutions are readily available with a simple, two-step shift in statistical perspective.

First, focus needs to be placed on predictive modeling (i.e., predicting new observations based on patterns found in previous obser-

vations) rather than explanatory modeling (i.e., providing evidence for a hypothesis with significant effects in a given direction). This removes the burden of using existing theory (which may or may not exist) to dictate what should or should not be specified in a given statistical model (Shmueli, 2010). The second step involves utilizing algorithmic methods that assume the data generative mechanism is unknown, rather than adopting a method that relies on the assumption that the data were generated from a given stochastic model (Breiman, 2001b). In other words, instead of assuming that the relation between an outcome and a predictor is a specific function of the given predictor value, its estimated parameter, and random noise, an algorithm treats the functional relation between the predictor and the outcome as unknown and something to be estimated. Algorithmic methods remove the burden of specifying complex function forms, such as nonlinear relations or higher order interactions, when attempting to build predictive data models.

Through this shift of perspective, it becomes possible to understand and adopt a popular data mining algorithm known as recursive partitioning to identify subsets of variables that are most related to a given outcome, and what kind of statistical relation it might be. This framework produces a set of binary decision rules based on covariate values in an attempt to create a set of subgroups, or nodes, which are homogeneous with respect to a given outcome (McArdle, 2013). This is particularly relevant in the context of multilevel data, where cases or observations are nested (e.g., children nested within classrooms). In these situations, using traditional methods, such as linear regression, can result in an increased chance of detecting significant effects due to the broken statistical assumption of independence (Peugh, 2010). The general recursive partitioning framework, on the other hand, makes no such assumptions, indicating that this method could extend to multilevel data structures with little added complications. And yet despite its potential utility in the social sciences, the implementation of these algorithms in the face of complex, multilevel data structures, is not well understood (McArdle, 2013).

The purpose of this dissertation is to determine whether recursive partitioning is a feasible tool to conduct exploratory data analysis in the presence of multilevel data, and, if so, which underlying algorithm yields the best results. This dissertation is structured into seven chapters. Chapter 2 provides a foundational overview of the two popular algorithms under investigation, classification and regression trees (CART) and conditional inference trees (CTREE), as well as their ensemble method (i.e., forest) counterparts. This section also includes a brief overview of statistical terminology relevant to understanding the application of predictive models, such as bias, variance, and model validation. Chapter 3 covers previous research extending recursive partitioning frameworks to more advanced applications rel-

evant to complex, multilevel data structures. Additionally, this chapter provides a set of unanswered questions when considering the application of both CART and CTREE to multilevel data and includes a set of hypotheses based on previous research and statistical theory.

The next three chapters provide the results for the three main phases of this dissertation: the simulation phase, the application phase, and the dissemination phase. Chapter 4, the simulation phase, provides an overview of the simulation design and discusses the results in the context of creating a set of best practices for researchers interested in using these techniques on multilevel data. Chapter 5, the application phase, uses these best practices to analyze three applied datasets in order to gain a better understanding for how the simulation results might generalize to real data. Chapter 6, the dissemination phase, gives an overview of an R package created to house code for researchers to use in order to extract meaningful information from these recursive partitioning models. Additionally, this chapter will include feedback received during a workshop given on these techniques, and include a discussion on how applied researchers felt these methods could be useful in their own work. Finally, Chapter 7 will be a general discussion tying together the overarching themes of this project to discuss the answers to the original research questions, as well as the limitations and future directions of this dissertation project.

2

RECURSIVE PARTITIONING AND ENSEMBLE METHODS

Recursive partitioning is a non-parametric, data-driven statistical algorithm that iteratively searches a given predictor space to identify potential splits, thus segmenting the space into distinct, rectangular subsections. Then, a simple model, most often a constant, is fit in each one (Hastie, Tibshirani, & Friedman, 2009). In the case of a continuous outcome, this recursive partitioning procedure is referred to as a *regression tree*, and the constant reflects the mean value of all observations in a given subsection. Otherwise, if the outcome is categorical, it is referred to as a *classification tree*, and the constant is assigned to be the specific category with the highest frequency over all observations in the particular subsection. As such, recursive partitioning is not limited to a binary outcome, but can actually model an outcome with K potential classes.

This framework was initially formulated through the seminal work of Morgan and Sonquist (1963), who proposed focusing on predictive accuracy in large survey data as a means to relax additive assumptions commonly made at that time in order to detect complex relations among predictor variables; a process they referred to as Automatic Interaction Detection. This framework was further solidified by both Breiman, Friedman, Stone, and Olshen (1984) and Quinlan (1986), who created algorithms seeking to improve on this original idea. Initially, recursive partitioning was met with much disdain by the statistics community due to its poor generalizability, especially when inappropriately applied to small datasets with large noise-to-signal ratios (Hastie & Tibshirani, 2013). However, this recently changed in the past decade or two, when the work of Breiman et al. (1984) started to become widely adopted as a useful tool in predictive modeling. Still, this method is relatively unknown to those in the social sciences (McArdle, 2013).

Given the immense popularity of the recursive partitioning framework, it is no surprise that many algorithms exist, each with their own advantages and disadvantages. This dissertation focuses on two of the most widely used recursive partitioning algorithms: classification and regression trees (CART; Breiman et al., 1984) and conditional inference trees (CTREE; Hothorn, Hornik, & Zeileis, 2006). These two algorithms, in addition to other important concepts inherent to predictive modeling, are described below. To assist in this overview and to highlight how the methods are used in practice, a dataset examining the relation between graduation rates and various statistics for US

Colleges (N = 777) from the 1995 issue of US News and World Report will be used. This dataset is freely available from the ISLR package (James, Witten, Hastie, & Tibshirani, 2013b) in the R software environment (R Core Team, 2014), and contains 17 variables that can be used as predictors for graduation rate, such as whether a university is private or public, the acceptance rate, and the out-of-state tuition cost.

2.1 CLASSIFICATION AND REGRESSION TREES

Originally proposed by Breiman et al. (1984), CART is easily the most popular and widely used recursive partitioning algorithm, being cited almost 28,000 times according to Google Scholar[1]. When initially creating a decision tree, the algorithm consists of essentially four steps (James, Witten, Hastie, & Tibshirani, 2013a). First, all predictor variables are initially considered for potential splits in a *greedy*, *top-down* manner. The splitting process is greedy, because each step searches for the best possible split at that time, rather than considering previous or future steps. It is top-down, because it begins with all observations belonging to the same group, or node, and then subsequently conducts splits further down the tree after initial splits have been made. Second, the best potential split is identified by some criterion, which is taken to be the residual sums of squares (RSS) in the case of a continuous outcome. Following the notation of Hastie et al. (2009), suppose we have a predictor variable x_j with a given split point s, which splits the predictor space into two regions: $R_1(j,s) = \{x \mid x_j < s\}$ and $R_2(j,s) = \{x \mid x_j \geqslant s\}$. The algorithm searches for a particular j and s that minimizes the following equation:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{1}$$

where $\hat{y}_{R_k}$ is the mean of the outcome within the kth subsection and $\sum_{i:x_i \in R_k(j,s)} (y_i - \hat{y}_{R_k})^2$ is the RSS over all participants in the kth subsection.

Once the best split is identified, the data is split on this threshold, creating two new subsections, or *child nodes*. The same procedure outlined above is then performed separately on each of these nodes, and this process is repeated until some stopping criterion is reached. For example, the algorithm might require a minimum number of observations to belong to a given node, regardless of whether another split will result in a reduction of the RSS. Finally, the given node becomes a *terminal node* once no further splits can be made within that node. When all nodes can no longer be split any further due to these stopping criteria, the algorithm terminates.

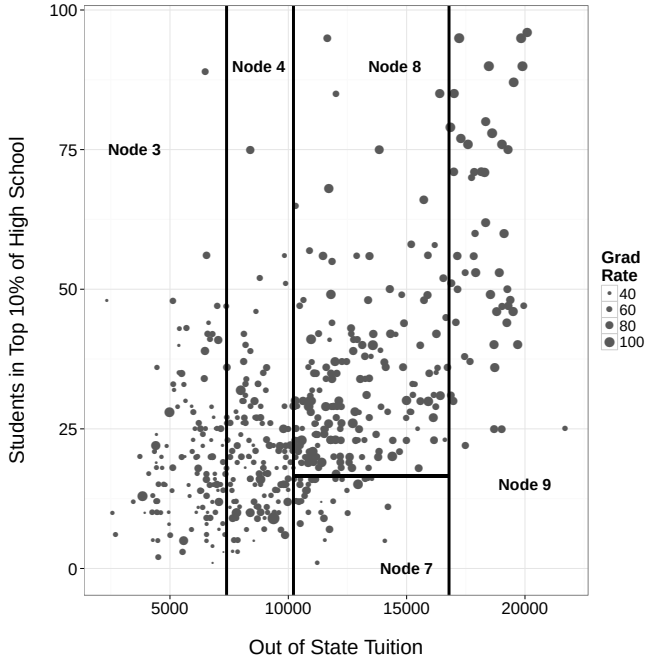
1 As of May, 2015

See Figure 1 for an example decision tree might look in the college graduation rate dataset. The corresponding decision tree first splits on the out-of-state tuition variable, which is a variable reflecting the annual tuition cost a student has to pay to attend the institution when they live in a different state (range: $2340 - $21700). This decision creates a vertical line in two-dimensional space, splitting the data into two subsections at approximately $10,000. Both of these nodes are again split by the out-of-state tuition variable, resulting in two more lines being drawn and creating four subsections in total. Finally, one node is further split by the percentage of students at an institution who were in the top ten percent of their high school class, resulting in a horizontal line creating a fifth subsection. Note that only two variables were used in the splitting process to maintain interpretability with the corresponding visualization.

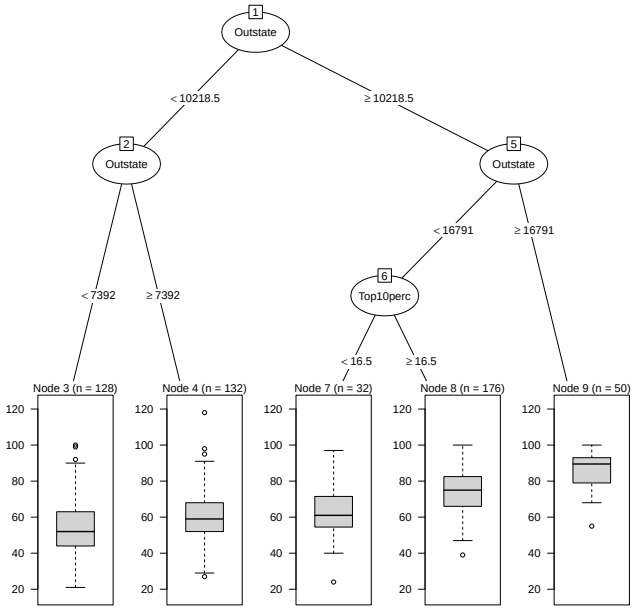
2.1.1 *Understanding the bias-variance tradeoff*

In the previous example, we see the last split results in a subsection with a small sample size (N = 32). When growing a decision tree on an initial dataset, referred to as the *training set*, creating a tree with more depth and complexity will yield excellent performance. However, because of the subtleties in the data at hand, the tree is likely to have poor performance generalizing to independent data. In other words, the tree algorithm is *overfitting* the data. This highlights the importance of validating a decision tree (or any predictive model for that matter) on a separate dataset, referred to as the *test set*, to get a more accurate estimation of its generalizability.

The underlying reason overfitting occurs is due to the *bias-variance tradeoff*. Conceptually, bias refers to the difference between a true value and the average prediction of said value over many sample datasets drawn from the same population. Variance, on the other hand, refers to how much variability exists in your predictions for a given data point. Models of low complexity typically have high bias and lower variance, while models of high complexity typically have low bias and high variance. Overfitting occurs when your model has too much variance due to being overly complex, resulting in predictions that generalize poorly to new data (Hastie et al., 2009). Thus, this bias-variance tradeoff is a balance between trying to identify a model with low bias that is still parsimonious enough to perform well on new data. Refer to Figure 2 to see how predictive performance varies as a function of model complexity between a training set and a test set in the graduation rate data. As the predictive model becomes more complex (i.e., more splits in the case of a decision tree), the estimated error always improves in the training set, but only improves to a certain extent in the test set. In this case, there is evidence for



(a)



(b)

Figure 1: *(a) shows how college graduation rates vary as a function of out-of-state tuition and percentage of students in the top 10% of their high school class. These partitions are created by a set of binary decision rules, shown in (b).*

overfitting when the decision tree contains more than three or four splits.

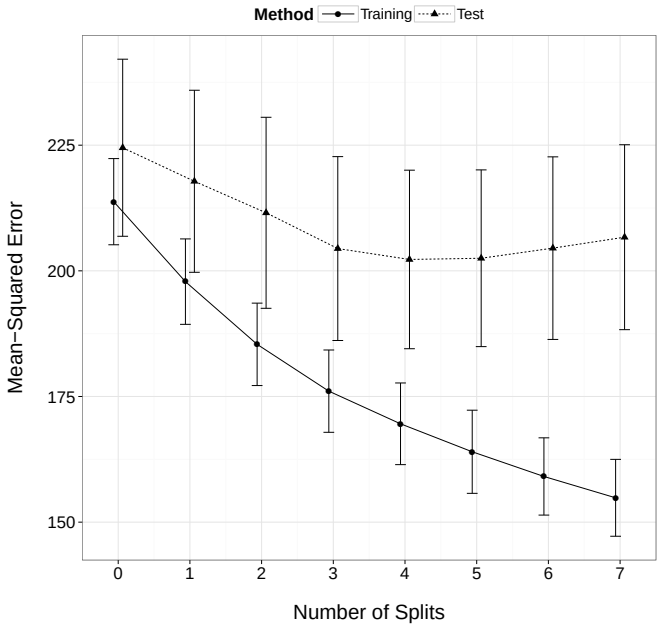


Figure 2: *The bias-variance tradeoff found in the graduate rate data, depicting how the error changes as a function of model complexity. The data was split into training and test sets 100 times to estimate precision, and the bars represent 1 standard error. Note that this figure was inspired by (James et al., 2013a).*

The bias-variance tradeoff can also be understood mathematically as a decomposition of the residual term in a given statistical model. Given a training set consisting of data points $x_1, \ldots, x_n$, each with a corresponding y_i, we can assume there exists some true functional relationship, $y_i = f(x_i) + \epsilon$, where ϵ is a noise term with a mean of zero and variance σ^2. By choosing some statistical method, say linear regression, we get an estimate of $f(x)$, namely, $\hat{f}(x)$. The total squared error of this estimated function relation is

$$\begin{aligned} \mathbb{E}[(y - \hat{f}(x))^2] &= \left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2 + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2] + \mathbb{E}[\epsilon^2] \\ &= \mathtt{Bias}(\hat{f}(x))^2 + \mathtt{Variance}(\hat{f}(x)) + \sigma^2 \end{aligned} \tag{2}$$

2.1.2 *Pruning decision trees with cross-validation*

Because decision trees that are too complex typically yield poor generalization performance, an additional step must be introduced into the process of building and evaluating a decision tree. How this issue is typically handled is with a procedure known as *pruning*. That is, trees are initially grown out to their maximum depth (i.e., maximum

complexity) with the algorithm outlined above. Then, the depth of the tree is reduced based on a given cost function. For any subtree T that is a subset of a decision tree grown to full length, define $|T|$ as the number of terminal nodes in T, and $\alpha \geqslant 0$ as the complexity parameter. The updated cost-complexity measure, $RSS_\alpha(T)$, is

$$RSS_\alpha(T) = RSS + \alpha|T| \tag{3}$$

Here, the complexity parameter can be tuned to control how complex the resulting tree is. If $\alpha = 0$, no pruning will be done. When this value increases, however, the depth of the tree will get smaller.

The cost complexity parameter only has a finite number of values, because there is only a finite number of subtrees within a given tree. Once all the alpha values corresponding to unique subtrees have been collected, the value correponding to the best performance is selected with a procedure known as *k-fold cross-validation*. This procedure first divides the training set into K equal-size partitions, or folds. First, a full tree is grown on all but the kth fold. The mean squared error is then extracted from the left-out kth fold as a function of the complexity parameter. This procedure is repeated K times, so every partition has an opportunity to act as a test set. The complexity parameter value that leads to the lowest average error across the K estimates is chosen for the creation of the final tree to be used for future predictions. Note that setting K as five or ten yields a good approximation to the true test error in many applications, so these values for K are typically chosen (Hastie et al., 2009). Another rule of thumb for selecting the value of the complexity parameter is known as the *1-SE rule* (Breiman et al., 1984). Because the cross-validation estimate of the complexity parameter results in a distribution of error estimates, the value of the complexity parameter is selected such that the simplest tree within one standard error of the best tree is chosen.

Figure 3 shows how the 1-SE rule is applied in practice for the college graduation rate dataset. Recall that Figure 2 indicated that three or four splits should be made, while the 1-SE rule indicates only one split should be made. This highlights a feature of the 1-SE rule, in that it typically selects more parsimonious models, which may or may not yield better generalization performance. Thus, the result of the 1-SE rule should be compared with other evidence, like Figure 2, to help determine how best to prune a decision tree.

2.1.3 *CART with a categorical outcome*

While not discussed in the introduction to CART above, the algorithm can easily be extended to a categorical outcome with K potential classes by simply modifying the RSS criteria for both the splitting and pruning. More specifically, three methods exist as a measure of

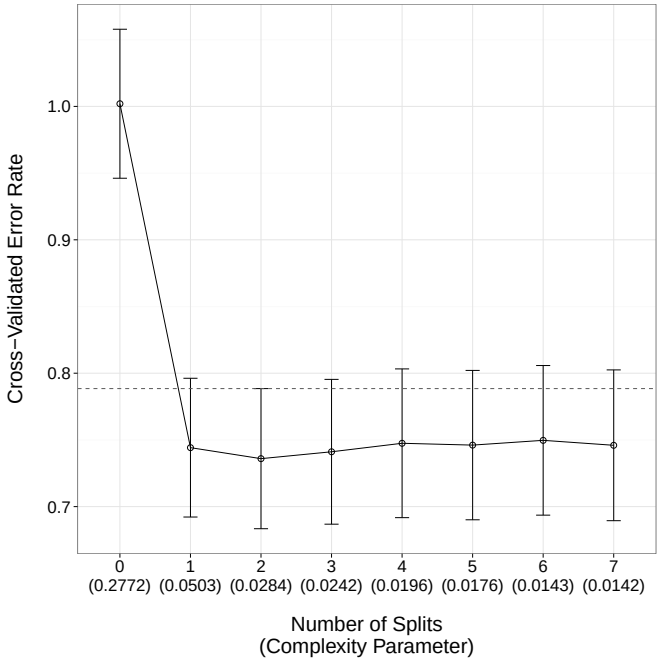


Figure 3: *How the 10-fold cross-validation error changes as a function of model complexity in the college graduation rate dataset. The 1-SE rule is set with the dotted line, indicating that this rule of thumb would select one split. Just selecting the tree that yields the minimum cross-validated error would select a tree with two splits.*

node purity in the case of a categorical outcome: classification error, the Gini index, and cross-entropy. Classification error is the simplest of the three, and is the fraction of observations in the training set that do not belong to the most common class

$$\text{Classification error} = 1 - \max_{k}(\hat{p}_{jk}) \tag{4}$$

where $\hat{p}_{jk}$ corresponds to the proportion of observations in the training set in the jth subsection that are from the kth class. While the classification error rate only includes the most common class, the Gini index reflects a measure of total variance across all possible classes

$$\text{Gini} = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk}) \tag{5}$$

Finally, cross-entropy is numerically similar to the Gini index, and is defined as

$$\text{Entropy} = -\sum_{k=1}^{K} \hat{p}_{jk}\log(\hat{p}_{jk}) \tag{6}$$

The relation between these three measures can be seen in Figure 4. When the probability of belonging to a class in a given node becomes more certain (i.e., closer to 0 or 1), the values of all three measures get closer to zero. Otherwise, when the probability of an observation belonging to a specific class gets more uncertain (i.e., closer to 0.5), the value of these measures increase. In other words, these measures calculate how homogenous a given node is, and the goal of the recursive partitioning algorithm is to minimize this value over all nodes in the splitting process. Note that either the Gini index or cross-entropy are typically favored in the splitting process, as they are more sensitive to node purity when compared with the misclassification rate (James et al., 2013a). If the probabilities of a given class are small, cross-entropy is likely a better choice compared with the Gini index, because of the log in the calculation. However, research has found no conclusive evidence regarding whether the Gini index or cross-entropy yields better performance in a general sense (Raileanu & Stoffel, 2004).

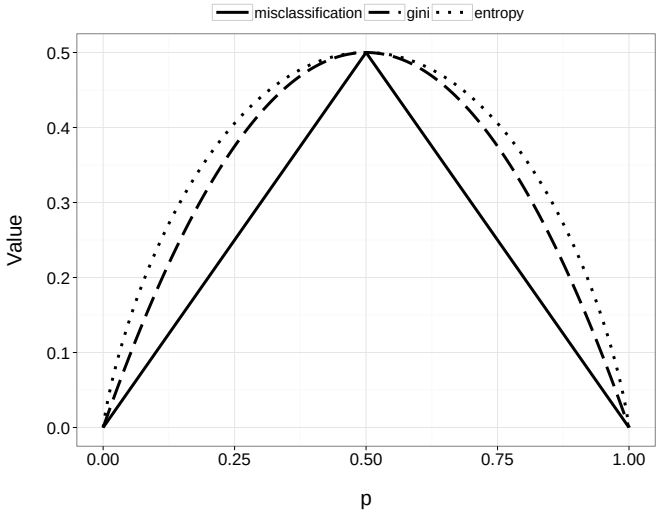


Figure 4: *The relation between classification error, the Gini index, and cross-entropy at different probability values for belonging to a given class when two classes are present. Image adapted from (Hastie et al., 2009).*

2.1.4 *Pros and cons of CART*

CART is a method that can efficiently search a parameter space, capturing potential nonlinear relations as well as higher-order interactions without explicit model specification by the researcher. It can also handle both continuous or categorical variables as outcomes by simply changing the underlying measure of node purity. Finally, and perhaps the most relevant to this dissertation, is that the resulting pruned tree is fairly easy to understand and explain to others, even if

those individuals are not familiar with the technique or lack a formal statistical background.

Despite all these advantages, CART is not without drawbacks. First, like all simple decision trees, they generally do not yield good predictive performance when compared to other regression-based techniques (James et al., 2013a). Pruning can also add a layer of researcher subjectivity in deciding how complex or simple a given tree is. Finally, because the splitting procedure considers all possible splits within all possible variables simultaneously, variables with more potential splits are more likely to be chosen to have a potential split point purely due to chance when compared with variables with less potential splits (Loh & Shih, 1997; Quinlan & Cameron-Jones, 1995). For example, there are $N - 1$ potential splits for a continuous variable (assuming no identical values among observations) and $2^{K-1} - 1$ potential splits for categorical variables with K categories. Assuming a sample size of 100, a continuous variable with no repetitive numbers or a categorical variable with eight classes will have 99 and 127 potential splits respectively, while a 5-point Likert scale item or a dichotomous variable will have substantially less (4 and 1, respectively).

2.2 CONDITIONAL INFERENCE TREES

To mitigate the issue of pruning and the biased splitting procedure found in many recursive partitioning algorithms, Hothorn, Hornik, and Zeileis (2006) proposed conditional inference trees (CTREE). While conceptually similar to CART, CTREE is based on a well-defined theory of permutation tests developed by Strasser and Weber (1999). Like CART, CTREE also consists of four main steps, which will first be explained conceptually. First, a global null hypothesis of independence between Y and all covariates X_j is tested. For a common example, if all covariates and the outcome are continuous, a Pearson's correlation coefficient with a corresponding p-value is calculated for each variable's association with the outcome. If no p-value is below the pre-selected alpha level after accounting for multiple significance tests, the global null hypothesis is not rejected and the algorithm terminates. Otherwise, the covariate with the stongest association (i.e., p-value) with the outcome is selected for splitting. The best split within this covariate is selected, and the training set is partitioned on this value. Finally, these steps are iteratively repeated until the global null hypothesis can no longer be rejected in all subsections.

These four steps can also be explained mathematically following notation from Hothorn, Hornik, and Zeileis (2006) and Molnar (2013).

Step 1: The Global Null Hypothesis Test

First, in a given sample, assume there exists a response Y and a covariate vector $X = (X_1, \ldots, X_p)$. Let $D(Y|X)$ represent the conditional distribution of response Y given the covariates X, and L_n represent the training set, such that $L_n = [Y, X_1, \ldots, X_p]$. The global null hypothesis is then defined as $H_0 = \bigcap_{j=1}^{p} H_0^j$, where $H_0^j : D(Y|X_j) = D(Y)$. This is tested with the following test statistic.

Let g_j be a transformation of the covariate X_j, such that $g_j : X_j \to \mathbb{R}^{p_j}$, where p_j is the number of dimensions of covariate X_j. The *influence function*, h, performs a similar transformation on the outcome, such that $h : y \to \mathbb{R}^q$. Also, let w correspond to a set of weights indicating whether a given observation is in a node or not (i.e., set at 1 or 0). The test statistic for every covariate j can be defined as[2]

$$T_j(L_n, w) = vec\left(\sum_{i=1}^{n} w_i g_j(X_{j,i}) h(Y_i)^T\right) \in \mathbb{R}^{p_j q} \tag{7}$$

where the resulting $p_j \times q$ matrix is converted into a $p_j q$ column vector using the *vec* operator. Note that this final conversion is not a necessity, but is done for computational simplicity.

Both the transformation and influence function are chosen depending on the scales of X_j and Y, respectively. An outline for selecting both g and h can be found in Hothorn, Hornik, and Zeileis (2006). In the frequent case of a numeric variable for both covariate and outcome, both functions can be the identity, indicating that no transformation is made. In the case of a categorical variable, the transformation for said variable can map the covariate into a variable with the number of dimensions equal to the number of categories (i.e., two dimensions in the binary case, with $(1, 0)^T$ for one condition and $(0, 1)^T$ for the other).

Because the distribution of T_j under the null hypothesis is unknown for most instances, it must be estimated via a *permutation test* framework. That is, a null distribution is contructed by fixing the covariate values and permuting the outcome to destroy the relationship between the covariates and the outcome. The derivation of the conditional expectation μ_j and covariance Σ_j under the null can be found in Hothorn, Hornik, and Zeileis (2006), and are not included here for brevity.

With the conditional expectation and covariance, $T \in \mathbb{R}^{pq}$ can be standardized. Let t be the value of T, μ be the expected value of T and Σ_{kk} is the kth diagonal entry of the covariance matrix under the

2 Note that this notation differs slightly from Hothorn, Hornik, and Zeileis (2006), where $h(Y_i)$ is written as $h(Y_i, (Y_1, \ldots, Y_n))^T$. This was removed for simplicity.

null distribution. Using these, the maximal entry of T after standardization is chosen for a test statistic c:

$$c(t, \mu, \Sigma) = \max_{k=1,\ldots,pq} \left| \frac{(t-\mu)_k}{\sqrt{\Sigma_{kk}}} \right| \tag{8}$$

The test statistic for each variable, $c(T_j(L_n, w), \mu_j, \Sigma_j)$, is then converted to a p-value scale in order to test the global null hypothesis and for the purpose of variable selection. To reiterate the most recent steps incorporating the permutation test, all observations in a given subsection are randomly permuted, and c is calculated for each covariate. The resulting permuted distribution for a given variable then corresponds to the null distribution assuming no relationship between that variable and the outcome, and a p-value can be calculated as the proportion of the null test statistic distribution that is larger in magnitude than the observed test statistic. The p-values for each variable are corrected for a global test using a multiple comparisons procedure (e.g., Bonferroni), and then compared to the preset α level which is typically (and arbitrarily) defined to be 0.05. If the global test is not rejected, the algorithm makes this node a terminal node.

Step 2: Variable Selection

If the global test is rejected, then the variable with the lowest p-value calculated in Step 1 is selected for splitting.

Step 3: Split Point Selection

Once the variable has been selected, splits are determined using a special case of the test statistic formula defined in Equation 7. For a selected variable, X_{j^*}, the equation is

$$T_{j^*}^A(L_n, w) = vec\left(\sum_{i=1}^{n} w_i I(X_{j^*,i} \in A) h(Y_i)^T \right) \in \mathbb{R}^q \tag{9}$$

where A is a possible partition of the current observations and

$$I(X_{j,i} \in A) = \begin{cases} 1 & : X_{j,i} \in A \\ -1 & : X_{j,i} \notin A \end{cases} \tag{10}$$

Note that the values are coded as 1 and -1 so if a given covariate is independent of the outcome, the corresponding values will typically cancel and yield a value for $T_{j^*}^A$ that is close to zero.

As defined, this sequence will test all possible subsets of A. To reduce the amount of computation, possible subsets for continuous

and ordinal variables are restricted to maintain their order among observations. An additional contraint can be made by requiring a minimum node size for all groups within a given subset. Finally, a partition, A^*, is identified as the one which maximizes the standardized test statistic using Equation 8 with the proposed new split A. That is,

$$A^* = \underset{A}{\operatorname{argmax}}\, c(t_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A) \tag{11}$$

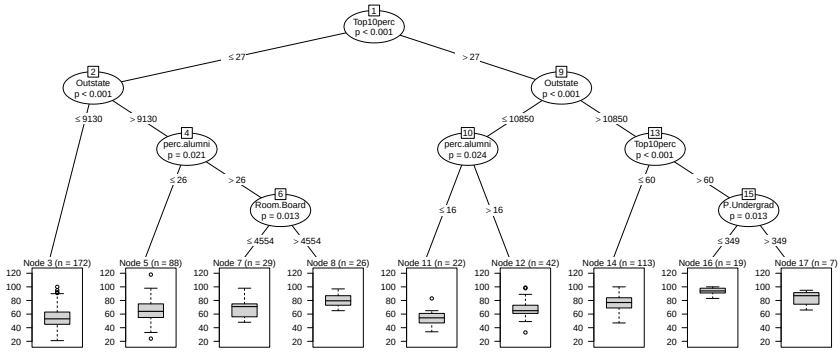
Step 4: Repeat

The weights within each node are re-normalized, and these steps are repeated until the global null hypothesis is no longer rejected in each subsection.

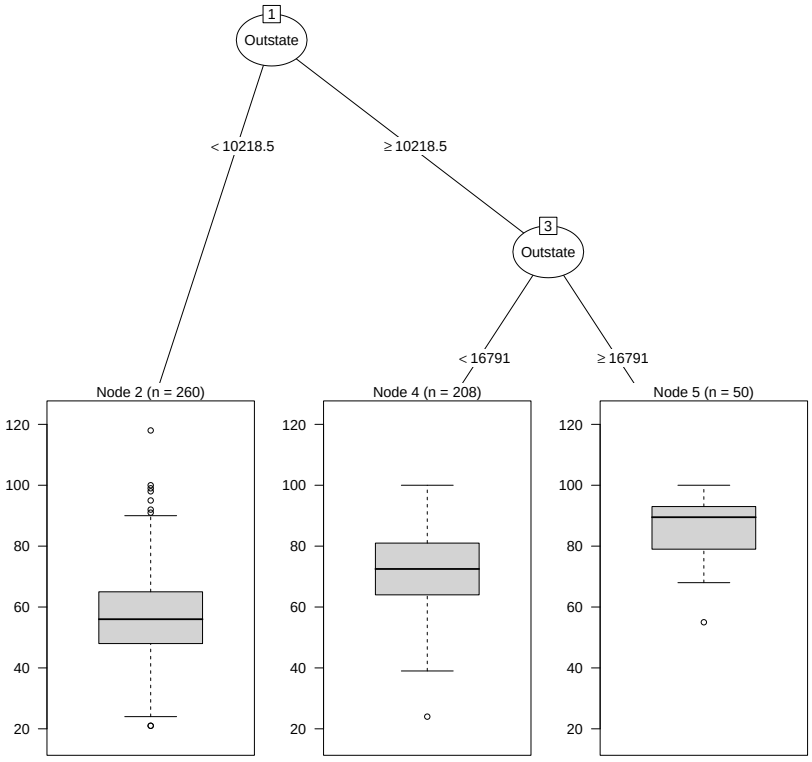
Clearly, this algorithm is very similar to CART in its basic iterative partitioning structure. Similarly, conditional inference trees can be pruned by altering the significance level required in the splitting process, which can be estimated via a cross-validation procedure (Hothorn, Hornik, & Zeileis, 2006). Refer to Figure 5 for a visual comparison between the final CTREE and CART models using the graduation rate. These two methods often yield similar predictive performance despite having different tree structures.

2.2.1 *Pros and cons of conditional inference trees*

Conditional inference trees house all the benefits of decision trees while simultaneously alleviating the issues of both pruning and biased variable selection found in CART. And yet, despite this lack of bias in the splitting procedure, CTREE and CART typically perform similarly with regard to predictive accuracy (Hothorn, Hornik, & Zeileis, 2006; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). One downside to this method is that, because it incorporates a permutation framework, the algorithm is much more computationally expensive compared to CART. Thus, if the goal is to simply achieve good predictive accuracy in a tree framework, CART is typically a much better choice. Another major con with the conditional inference framework is that it assumes the data adhere to the traditional assumptions of independence. This potential issue with regard to multilevel data will be discussed further in Chapter 3. Finally, like all decision trees, CTREE is a predictive method that is often outperformed by regression techniques. There are two main reasons for this. First, trees are predictive methods that possess high variance. Oftentimes, a small shift in the training set (e.g., sampling the training set more than once) will result in a completely different tree structure



(a)



(b)

Figure 5: *A comparison of a conditional inference tree (a) and a pruned CART tree (b). Note that the conditional inference tree contains p-values refering to the significance of each split. Despite a different tree structure, both trees yield similar predictive performance.*

being detected (Hastie et al., 2009). Second, smooth relationships often naturally occur in datasets, and can be better captured with the underlying additive structure found in regression methods compared to the binary, piecewise splits underlying decision trees.

2.3 RANDOM FORESTS

As mentioned previously, while trees provide intuitive split criteria for a given dataset, these decision rules suffer in generalizability performance due to having high variance. Breiman (1996) proposed a solution to this issue by repeatedly taking a bootstrap sample of the training set to create many decision trees, and then aggregating the results of all the trees together to create the final predictions. This technique, referred to as *bagging* (short for bootstrap aggregating), leads to better predictive performance due to the fact that trees grown to full length are predictors that have low bias and high variance. Essentially, by aggregating their predictions into an ensemble, this new method maintains the low bias found in single decision trees, but also has reduced variance by replacing the hard, piecewise fits found in decision trees with smoothed relations (Breiman, 1996; Bühlmann & Yu, 2002; Strobl et al., 2009).

One more improvement to this idea was made by Breiman (2001a). Due to the greediness of the recursive partitioning algorithm, a bagged ensemble of trees is typically dominated by one or two variables that are always used in the first split. This ignores the possibility of the existence of a tree with better predictive performance that contains a first split that is suboptimal. To give these lesser predictors a better chance to be incorporated into the splitting procedure, Breiman (2001a) introduced *random forests*, which first grew decision trees using a random subset of possible predictor variables, and then used bootstrap aggregation to create an ensemble. In this method, both the number of variables to select for each tree and the number of trees to grow in total are parameters that can be tuned to yield better predictions. Many statistical software packages have sensible defaults for these values that typically yield good performance, suggesting that a random forest is a good "off-the-shelf" method that does not require as much tuning when compared to other predictive models (Strobl et al., 2009).

2.3.1 *Out-of-bag samples*

In a bootstrap sample, the probability of an observation being selected is $1 - (1 - \frac{1}{n})^n$. As $n \to \infty$, this value converges to approximately 0.632. Thus, by nature of the bootstrap process, approximately 63% of the dataset will be used for training in each decision tree. This highlights a nice feature of random forests, in that about 37% of the

data, referred to as the out-of-bag (OOB) sample, are not used to train any given tree. Using the OOB samples to evaluate performance has been shown to be a good approximation of the error found in a separate test set (Hastie et al., 2009), and are typically used as a replacement for the cross-validation step found in training predictive models.

2.3.2 *Variable importance*

Because random forests include many trees, each with their own distinct set of decision rules, they inevitably become harder to interpret. However, both numerical and graphical methods do exist. One such method is known as *variable importance*, and is typically measured in a permutation framework. More specifically, once an ensemble method has been created, the OOB samples are used to estimate predictive performance. Then, the same data is permuted with respect to a given variable in order to break that variable's link with the outcome, and the predictive accuracy of the overall forest is measured again. Finally, variables are assigned a value that corresponds to the difference in the original prediction accuracy and the permuted prediction accuracy. If the difference is large, then this indicates that the particular variable plays a more important role in predicting the outcome. Otherwise, if the predictive accuracy does not change very much, then this variable does not play a large role in predicting the outcome. See Figure 6 for the variable importance plot for the college graduation rate example. This plot confirms what was found in the single decision tree: out-of-state tuition is the most important variable for predicting college graduation rate.

2.3.3 *Partial dependence plots*

While variable importance metrics are useful, the functional relations between the variables and the outcome remain hidden from view. One way to further investigate these relations is with *partial dependence plots*. These plots are graphical visualizations of the marginal effect of a given variable (or multiple variables) on an outcome. Typically, these are restricted to only one or two variables due to the limits of human perception, and thus may be misleading due to hidden higher-order interactions. Despite this, partial dependence plots can still be extremely useful for knowledge discovery in large datasets, especially when the random forest is dominated by lower-order interactions and main effects.

Following the notation of Hastie et al. (2009), partial dependence plots can be mathematically defined as follows. Suppose S is a subset of p predictor variables, such that $S \subset \{X_1, X_2, \ldots, X_p\}$. Let C be a complement to S, such that $S \cup C = \{X_1, X_2, \ldots, X_p\}$. The random forest

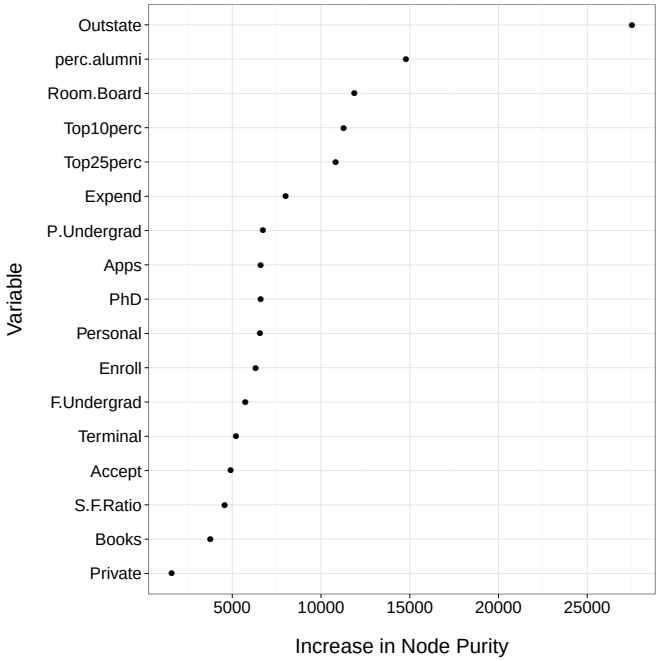


Figure 6: *Variable importance for a CART random forest, indicating the out-of-state tuition has the largest role in predicting graduation rate.*

predictor function, $f(X)$, will depend upon all p predictor variables. Thus, $f(X) = f(X_S, X_C)$. The partial dependence of the S predictors on the predictive function $f(X)$ is

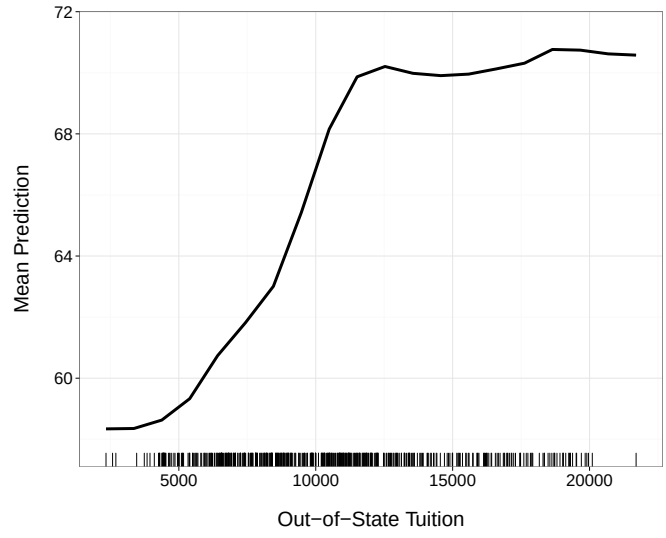
$$f_S(X_S) = \mathbb{E}_{X_C}[f(X_S, X_C)] \tag{12}$$

and can be estimated by

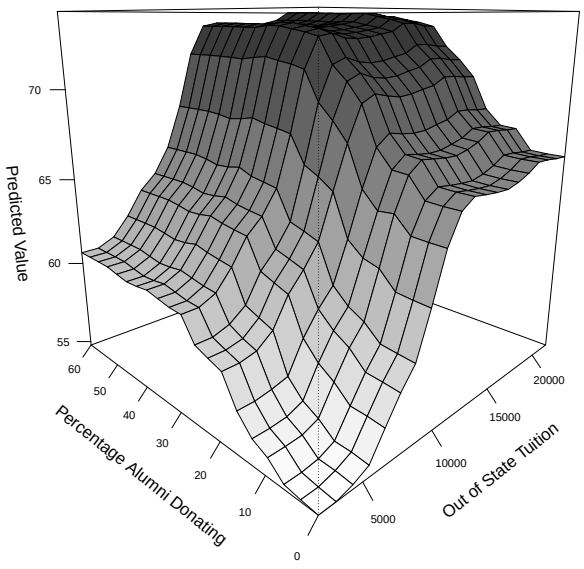
$$\bar{f}_S(X_S) = \frac{1}{N}\sum_{i=1}^{N}[f(X_S, X_{Ci})] \tag{13}$$

where $\{x_{C1}, x_{C2}, \ldots, x_{CN}\}$ are the values of X_C ocurring over all observations in the training data. In other words, in order to calculate the partial dependence of a given variable (or variables), the entire training set must be utilized for every set of joint values in X_S. As one can image, this can be quite computationally expensive when the dataset becomes large. See Figure 7 for examples of partial dependence plots for the college graduation rate dataset.

Because partial dependence plots are computationally inefficient for even moderate sample sizes (> 1000), they can be time-consuming tools to extract information from a forest model. This is especially problematic here, given that the focus of this dissertation is on exploratory data analysis. Thus, researchers will be interested in examining and comparing predicted functional relations for many individual variables simultaneously. Luckily, a simplification can be made in



(a)



(b)

Figure 7: *(a) shows the partial dependence plot for out-of-state tuition, revealing a somewhat linear relation with graduation rate. (b) shows the partial dependence plot between both out-of-state tuition and percent of alumni donating to the institution, resulting in a three-dimensional plot. Note that because both variables have consistent relations with graduation rate across the values of the other variable, there is no substantial evidence for a potential interaction.*

the calculation of a partial dependence plot that results in vastly improved computation time with minimal loss of statistical information. Rather than calculating predicted values by repeating the training set across all joint values in X_S, simply create a new dataset of one observation that consists of a measure of central tendency for continuous variables and the most-endorsed level for either categorical or ordinal variables. Then, repeat this observation across all joint values in X_S. In addition to being much faster, these predicted values are still easy to interpret as the change in a given variable while holding all other variables fixed at values that represent an "average" observation in the dataset. While these plots are typically referred to as a "poor-man's" partial dependence plot (Milborrow, 2014), they will be referred to as *predicted value plots* in this dissertation for clarity. See Figure 8 for an example of a predicted value plot for the college graduation rate dataset.

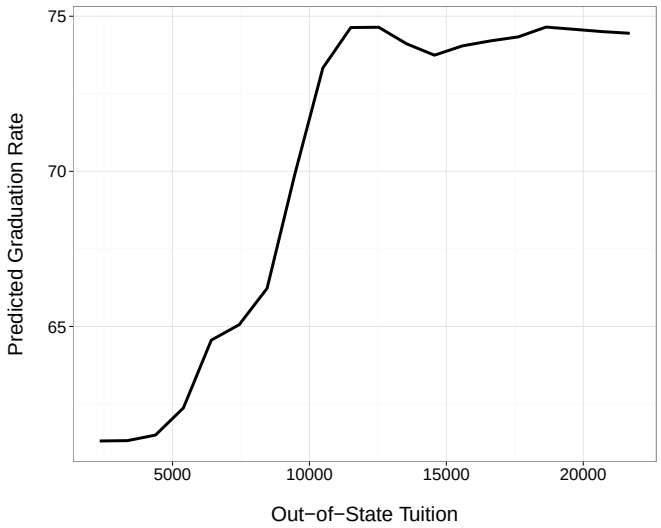


Figure 8: *Predicted value plot for out-of-state tuition variable predicting graduation rate. Despite being simpler and more computationally efficient to calculate, this plot shows a similar trend, although a higher intercept, to the true partial dependence plot in Figure 7.*

2.3.4 *Conditional inference forests*

In this overview so far, only a random forest with CART as the underlying partitioning algorithm has been used. In fact, random forests are easily generalizable to any recursive partitioning framework. In the case of conditional inference, for example, creating an ensemble of conditional inference trees is known as a conditional inference forest (CFOREST; Hothorn, Bühlmann, Dudoit, Molinaro, & Van Der Laan, 2006), and can be readily adapted with one slight alteration to the original forest algorithm. Because the framework of

Table 1: *Estimated prediction performance for the four methods.*

Method	Mean Performance (MSE)	SD Performance
CART	208.41	19.22
CTREE	205.28	17.44
CART forest	168.61	14.55
CFOREST	169.27	14.63

conditional inference assumes independence, taking a bootstrap sample results in a new dataset with repeated observations, thus breaking this assumption. In order to maintain an unbiased splitting procedure, this issue can solved by sampling 63.2% of the data without replacement instead of bootstrapping (i.e., subsampling). This improvement has shown to yield an unbiased splitting procedure with predictive performance on par with CART forests (Strobl et al., 2007). While the graphical results of the conditional inference forest will not be given for brevity, see Table 1 to compare the predictive performance of CART, CTREE, CART forests, and CFOREST applied to the college graduation rate data in 100 runs. Given the results of CART in Figure 2, the tuning parameter was selected via the minimum cross-validated error rather than the 1-SE rule. Note that the performance for both trees and forests are similar between the two methods of CART and conditional inference. While this is generally true, the underlying decision tree structure will be somewhat different given the biased nature of the CART algorithm (Hothorn, Hornik, & Zeileis, 2006; Strobl et al., 2007). In this specific example, the top 5 predictors and their relative order (with the exception of one variable) between both random forest methods were identical.

2.3.5 *Pros and cons of random forests*

As was just shown, the major benefit to creating an ensemble of decision trees with a random forest algorithm is being able to efficiently search a large parameter space and create a predictive model that is vastly superior to a single decision tree. Another added benefit is that forests are created on unpruned trees, removing the subjectivity inherently found in pruning decisions. However, this benefit of improved performance does come at the cost of reduced interpretability. While both variable importance and partial dependence plots offer ways to view the underlying structure of decision trees, these methods still mask the potential higher-order interactions underlying a given forest that are impossible to visualize. Additionally, forest methods are more computationally expensive to perform when compared to a single decision tree. This is especially true of CFOREST, which requires a more expensive permutation test framework. For example, a recent

application of random forests on 876 observations with 44 variables reported 4.82 seconds for CFOREST, while CART forests only took 0.24 seconds (Strobl et al., 2007). While this effect can be somewhat mitigated on larger datasets by the fact that forests are trivially parallelizable, this characteristic of longer computation time can be perceived as a nuisance, especially with extremely large datasets.

2.4 HANDLING MISSING DATA

Missing data are an all-too-common occurrence in many areas of the social sciences, especially in education. Fortunately, recursive partitioning has a few different ways to elegantly handle missing data. The first, surrogate splits, are typically used for decision trees and are handled internally in the algorithm itself. Every time a split is made in the procedure, the algorithm automatically searches for additional variables that mimic the behavior of this main split, thus acting as a surrogate to the original variable in the case of a missing observation. These splits are derived in the exact same way as the original partitioning algorithm (Hothorn, Hornik, & Zeileis, 2006; Therneau & Atkinson, 2014). More specifically, the algorithm searches for splits in other variables that best predicts membership into the two new classes created by the split in the original variable. A list of ranked variables is then created for each split in the final decision tree in case data are missing on multiple variables. While this approach seems simple, surrogate splits have been shown to yield performance similar to multiple imputation when applied to both CART and CTREE under missingness conditions that are completely at random (Hapfelmeier, Hothorn, & Ulm, 2012).

Unfortunately, while the application of the methodology of surrogate splits can extend directly to ensembles of trees for the purpose to calculating predictions, there is no clear way to estimate permuted variable importance in the traditional way. This is one reason as to why surrogate splits are not employed in many CART forest algorithms (e.g., Liaw & Wiener, 2002; Pedregosa et al., 2011). Instead, random forests often use imputation methods. For example, the original random forest algorithm employs an imputation method that iteratively builds random forests in order to impute missing values (Breiman, 2001a). To do this, missing values are first replaced by the median of that variable (in the continuous case) or the category with the highest prevalence (in the categorical case) of that particular variable. Then, a random forest is run, and the missing values are replaced based on their proximity to other observations. That is, when a missing value is present, observations are identified that typically appear in the same nodes as this missing value in many of the trees created during the random forests process. These observations are then used to impute the missing values. This procedure is repeated a few

times to iteratively improve on the final imputations. This method has also been shown to yield good performance even with higher rates of missingness (i.e., > 50%), though it has been noted that the OOB error estimate tends to be slightly optimistic with imputed values (Breiman, 2003). Note that both surrogate splits and this imputation method only work when covariates are missing. Observations with the outcome missing must be either removed or imputed with a separate process.

CFOREST utilizes surrogate splits and employs a new variable importance measure to allow the estimation of variable importance in the presence of missing data (Hapfelmeier, Hothorn, Ulm, & Strobl, 2014). This new measure is essentially identical to permuted variable importance, with one exception. Rather than randomly permuting a variable to break its link with the outcome, observations are randomly sent to either the left of the right child node of any split that includes the variable of interest. This still breaks a variable's link with the outcome, but now the calculation of variable importance can be accomplished as before. This new technique shows a number of desirable properties over MCAR, MAR, and NMAR data, such as not yielding artificially inflated variable importance values for variables with high rates of missingness, which can occur with complete case analysis (Hapfelmeier et al., 2014). This new variable importance metric is also used for CFOREST even if no missingness occurs, because it is more computationally efficient for the conditional inference algorithm. It is important to note that, by definition, these methods can only handle missingness in the predictors. To the author's best knowledge, there is no non-parametric way to handle missingness in both the predictors and the outcome at this time.

3

RECURSIVE PARTITIONING AND MULTILEVEL DATA

3.1 PREVIOUS RESEARCH

There is a small, but growing body of research that examines the application of decision trees to complex, multilevel data structures. Segal (1992) provides an initial foray into this topic by attempting to extend the logic of decision trees and covariate splits to longitudinal data. In this method, split functions for trajectories can be directed toward either the mean vector (e.g., an intercept and a slope term) or the covariance matrix. One large difficulty in this procedure is the handling of time-varying covariates. Because changes in the covariance structure are considered in the splitting function, potential splits can only be done at the participant level and not the observation level. Thus, time-varying covariates can only be included if they are aggregated to the participant-level of analysis as a low-order polynomial term. That is, using an intercept and slope to approximate the trajectory of the time-varying covariate, and then using these new variables as potential variables to split on (Segal, 1992). This is an important methodological consideration that will manifest itself throughout this discussion of previous research.

This method can also be utilized to identify representative curves for the purpose of exploratory data analysis (Segal, 1994). While just sampling curves can be problematic in a sense that potential non-representative trajectories are selected (i.e., outliers), this method can detect homogeneous sub-populations of trajectories in regard to both the outcome and a set of covariates. A similar approach can be made from an unsupervised perspective, where representative groups are selected based on trajectories only and not the covariates themselves (Martin & von Oertzen, 2015; Tucker, 1966).

Other methods focusing on applying recursive partitioning to longitudinal data employ improved algorithms that do not exhibit the selection bias commonly found in some recursive partitioning algorithms, such as CART. For example, Eo and Cho (2013) proposed a recursive partitioning algorithm based on GUIDE (Loh, 2002), which utilizes residual analysis to avoid selection bias in its splitting procedure and a cost-complexity stopping rule instead of cross-validation to save on computation time. Similar to Segal (1992), this method can only identify trends, not predict future responses. As such, it is limited in only being able to split on variables at the cluster level, not

the observation level. A similar method proposed by Loh, Zheng, et al. (2013) also utilized GUIDE for this purpose.

Note that these aforementioned methods are focusing on exploratory data analyses in a multilevel framework that is longitudinal in nature (i.e., repeated observations nested within participants). Emphasis is also being placed on predictive accuracy in a multilevel framework that is typically cross-sectional in nature. For example, Karpievitch, Hill, Leclerc, Dabney, and Almeida (2009) examined the classification performance of random forests in mass spectrometry-based studies, which often produces cluster-correlated data. Via simulation, they found that traditional random forest algorithms yielded good classification performance despite the presence of cluster-correlated data, except that the OOB error rate typically underestimated the actual error rate calculated on a separate test dataset. By extending the original random forest algorithm to instead re-sample at the cluster level rather than the observation level led to near identical median classification rates and variable selection accuracy when compared to the original random forest algorithm, but without the bias found in the OOB error rate. The conceptual background for why this occurs will be addressed later in this chapter.

Additional examples investigating the performance of resampling methods for random forests in cluster-correlated data are typically found in medical research. In one example, Adler, Brenning, Potapov, Schmid, and Lausen (2011) compared different resampling methods for random forests in the presence of paired organ data, and found that the decrease in performance at high correlations between organs can be reduced if a paired bootstrap is performed. It has also been reported that sampling one observation rather than resampling all observations within an individual can lead to better performance when classifying future observations of patients in a longitudinal framework (Adler, Potapov, & Lausen, 2011).

More recently, researchers are starting to use a random effects structure in tandem with decision trees to improve predictions. For example, both Sela and Simonoff (2012) and Hajjem, Bellavance, and Larocque (2011) independently proposed the same method to incorporate a given random effects structure in a recursive partitioning algorithm, called RE-EM trees and mixed-effects regression trees, respectively. Both approaches operate on the same algorithm, namely one that uses a variant of the EM algorithm to estimate a set of random effects (most commonly just random intercepts) to encompass an entire tree. While any underlying decision tree algorithm can be used, both authors adopted an approach using CART. Because these random effects can be used to alter predictions for each individual observation in the sample after filtering through the tree structure, both methods show increased in-sample predictive accuracy compared to both a traditional multilevel model with main effects or a decision

tree without this random effect structure. However, the predictive performance for observations nested within unobserved clusters were similar between a decision tree with and without a random effects structure, because random effects have a mean of zero (Hajjem et al., 2011; Sela & Simonoff, 2012).

Given the bias inherent in the splitting procedure of the CART algorithm, these methods are being built around conditional inference trees, resulting in unbiased variable selection with the added benefit of having a random effects structure to improve predictive accuracy. Despite the unbiased variable selection, the conditional inference tree with random effects still yields similar predictive accuracy when compared to a CART tree with random effects (Fu & Simonoff, 2015); a similar scenario to how these methods perform without a random effects structure (Strobl et al., 2007). Hajjem, Bellavance, and Larocque (2014) have also extended these trees to create an ensemble method referred to as mixed-effects random forests. This method first removes the estimated random effects in the model before performing bootstrapping to avoid employing a cluster bootstrap. It has substantial improvements over single decision trees both with or without random effects in terms of prediction accuracy.

Additional research is investigating decision trees in multilevel contexts from a model-based perspective. That is, rather than focus on prediction with a basic tree approach where each node just represents a different mean value (in the case of a regression tree) or class (in the case of a classification tree), an alternative is to have a statistical model in each node (Zeileis, Hothorn, & Hornik, 2008). For example, a linear regression tree searches the covariate space for potential splits that maximizes the difference in model parameters in an already specified model. Thus, it becomes possible to unite confirmatory and exploratory research and examine potential model misspecification in more detail (Kopf, Augustin, & Strobl, 2013). Recently, model-based decision trees have been extended to a structural equation modeling (SEM) framework, called SEM trees (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013). In this method, decision trees are combined with the flexibility of SEM, where SEMs are split based on a set of covariates, effectively creating a series of more homogenous multi-group models. This model-based recursive partitioning method also avoids selection bias with a two stage procedure found in Loh and Shih (1997), which first selects the best cut point for each variable and then selects the best split among these candidates. Like many tree methods, creating an ensemble (i.e., a SEM forest) helps to reduce bias and increase the stability of estimated relationships among variables (Prindle, Brandmaier, McArdle, & Lindenberger, 2014). Because this framework also estimates variance components, it is limited to only splitting on cluster level variables in the presence of clustered data.

3.2 CURRENT PROBLEMS

As shown in the previous section, extending the recursive partitioning framework to a variety of multilevel contexts has received much research attention, especially in the last few years. However, many questions remain unanswered with regard to the application of these methods in the social sciences. For one, this previous research is typically focused solely on predictive accuracy, rather than the data-driven identification of potential variables of interest in an exploratory context. Additionally, many of these previous examples attempt to extend recursive partitioning methods to very situational circumstances, inevitably making them less flexible. Moreover, the simulation studies typically mimic the data generating process found in scientific areas where these methods are common (e.g., genomics), which are often quite different than what is found in the social sciences more generally, and education specifically. For example, these studies typically involve predictor variables that have all been measured at the lowest level of analysis (i.e., level-1) and account for the cluster level (i.e., level-2) variation as more of a nuisance parameter. In the social sciences, however, it is common to include variables measured at different levels of the analysis. Finally, not one of the studies that was previously mentioned have actually examined the original non-parametric methods outlined in Chapter 2 in multilevel contexts commonly found in the social sciences. Because of this, the performance of these methods with regard to variable selection bias, variable importance accuracy, and missing data handling are not well understood. Below, an outline is given for potential issues that manifest themselves conceptually when considering the application of CART, CTREE, and forest methods to multilevel contexts commonly found in education research.

3.2.1 *Multilevel issues with CART*

Given the non-parametric nature of the CART algorithm, it would seem as though CART can be applied to multilevel contexts without much additional consideration. While this is mainly true, recall that the algorithm is inherently biased toward selecting variables with many potential split points. This effect could be compounded when considering whether the variable under consideration appears at the first level or the second level of analysis. With N observations nested within K clusters, the number of potential split points for a numeric variable (with no repeated values) at the first level will be $N-1$, while the number of splits for the same variable at the second level will be $K-1$. It is clear that if both of these variables have no relationship with the outcome, the variable measured at the lowest level will be selected more often purely due to chance.

The categorical case, however, will not be as affected by this methodological issue assuming the clusters have the same number of observations. For a simple example, assume we are interested in measuring the entropy of a node that has a sample size of 16 (four observations nested within four clusters), along with a variable that is dichotomous. Suppose there exists a 50 percent chance of belonging into one outcome group versus the other. Because the observations are balanced within each cluster, the entropy of both situations is identical, regardless of whether the group was assigned at the cluster level or the observation level.

3.2.2 *Multilevel issues with conditional inference*

As mentioned in Chapter 2, the conditional inference algorithm assumes that the data are independent. Thus, the splitting procedure will be more likely to select a split in the presence of cluster-correlated data (i.e., an increased chance of a false positive for splitting), resulting in a tree that is more likely to overfit due to being too complex. It is likely that while varying the level of the variation in the outcome due to the cluster-level (i.e., the intra-class correlation coefficient, or ICC) can result in some bias, the conditional inference procedure would be most affected by the presence of non-independence due to the inclusion of level-2 variables in the splitting procedure. This follows conceptually from traditional regression techniques, where standard error inflation most often occurs due to the presence of level-2 variables incorporated at the first level of analysis (Luke, 2004).

Despite this methodological issue, conditional inference trees may still be useful in certain situations. For example, the rate of alpha inflation on datasets with only level-1 variables may result in trees that are still unbiased with respect to their splitting criteria, but just overfit the data due to non-independence. Simply altering the complexity parameter with cross-validation could be a potential solution to this issue. Additionally, trees are typically grown to maximum depth when creating a random forest. Conditional inference forests might still yield good predictive performance in the presence of non-independence, because conditional inference trees will have much more complexity in this case which is then removed in the aggregation step found in random forests. Regardless, many researchers might not realize the assumptions inherent to conditional inference trees, and inappropriately apply these methods in multilevel contexts without considering the potential consequences. Thus, it is important to identify situations where conditional inference is completely unreliable and where it might still be useful.

3.2.3 *Multilevel issues with forests*

In the random forest methodology, recall that approximately 37% of all observations in a sample will not be chosen in each bootstrap, referred to as the OOB sample. In a cluster-correlated sample, however, observations within a given cluster are more related to other observations within that cluster compared with observations in other clusters. Thus, exposing a tree to a given observation within a cluster actually informs the tree about other observations within that same cluster. This results in trees that are correlated with one another, yielding an OOB error estimate that is overly optimistic (Karpievitch et al., 2009). This is most problematic when the intra-class correlation (ICC) is high, something not commonly found in cross-sectional multilevel models in the social sciences. Previous research (i.e., Karpievitch et al., 2009) suggests that this methodological artifact should not be too problematic with smaller ICC values. In other areas where higher ICCs are common (e.g., mass spectrometry), other non-parametric bootstrap methods might provide a more reliable OOB error estimate. An alternative solution for a more accurate estimate of the test error without altering the re-sampling process underlying the algorithm would be to use cross-validation on the clusters rather than the observations, which has been shown to provide approximately unbiased estimates of test error in the presence of dependent data (Rice & Silverman, 1991; Segal, 1992).

However, note that the OOB samples are not only used as an estimate of test error, but also used to create the permuted variable importance measures. Thus, an overly optimistic OOB error estimate could lead to additional bias being introduced to variable importance measures. Again, because this issue is not substantial when ICC values are low, it only expected to lead to negligible bias. Regardless, this issue remains an open question, and will be investigated more thoroughly in this dissertation.

4

SIMULATION PHASE

With the quantitative issues outlined in Chapter 3, the behavior of CART, CTREE, CART forests and CFOREST all need to be better understood in multilevel contexts before they can be used as exploratory data analytic tools in education research. The goal of the simulation phase is to evaluate the performance of recursive partitioning methods in situations commonly found in education research and identify where these techniques may or may not be reliable. Naturally, given the inherent complexities of multilevel data and the many unanswered questions outlined in the previous section, these techniques will not be perfect. However, this does not mean that they will not be useful.

The structure of this chapter is as follows. First, more explicit details regarding the implementation of these analytic methods used in this comparative simulation are outlined. After that, the simulation conditions and evaluation criteria of proportion variation explained and variable importance will be explained. Finally, the results of the simulation will be presented and discussed.

4.1 STATISTICAL TECHNIQUES AND IMPLEMENTATION

Each dataset in this simulation was subjected to an automated implementation of four recursive partitioning methods and two multilevel regression methods described next. All simulations were conducted with R, version 3.1.1 (R Core Team, 2014).

4.1.1 *Classification and regression trees (CART)*

Decision trees using the CART algorithm were run using rpart, version 4.1-8 (Therneau & Atkinson, 2014) using cost-complexity pruning and choosing a tree depth based on the minimum error rather than the 1-SE rule. Missing data was handled internally by way of surrogate splits. Variable importance was extracted from the fitted rpart object, which follows Breiman et al. (1984) to calculate total importance of a given variable as the sum of its improvement in node purity in the case of a primary split and improvement multiplied by agreement with the primary splitting variable in the case of a surrogate split.

4.1.2 *Conditional inference trees (CTREE)*

Decision trees using conditional inference were run using party, version 1.0-20 (Hothorn, Hornik, & Zeileis, 2006) with no pruning, which is generally the approach taken for conditional inference methods on independent data structures (Hothorn, Hornik, & Zeileis, 2006). While conditional inference trees can handle missingness internally via surrogate splits, they do not do so by default. Additionally, it was of interest to see how a conditional inference tree would perform on missing data if the imputation was inappropriately based on a CART forest. Thus, missingness was handled using a single imputation procedure based on the proximity matrix extracted from a CART forest. Finally, no variable importance function exists for conditional inference trees, so one was manually created that calculates permuted importance for each variable.

4.1.3 *Random forests using classification and regression trees (CART forest)*

CART forests were run using randomForest, version 4.6-10 (Liaw & Wiener, 2002) with the traditional defaults of 500 trees grown to maximal depth and the square root of the total number of predictors (rounded down) as the number of variables in each tree. Missingness was handled using a single imputation procedure based on the proximity matrix extracted from a CART forest. Variable importance was extracted from the fitted model object, which is based in a standard permutation framework.

4.1.4 *Random forests using conditional inference trees (CFOREST)*

Conditional inference forests were run using party, version 1.0-20 (Strobl et al., 2007; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008), with the defaults of 500 trees grown to a given depth decided by a minimum criterion of 0.05 and the number of predictors used in each tree set to be five. Each tree was built using sub-sampling rather than bootstrap re-sampling as suggested by Strobl et al. (2007) for unbiased variable importance. Missingness was handled using a single imputation procedure based on the proximity matrix extracted from a CART forest, in line with the secondary research question outlined in the conditional inference trees section above. Variable importance was extracted from the fitted model object, which is based in a permutation framework as described by Hapfelmeier et al. (2014).

4.1.5 *Multilevel regression*

Multilevel regression models were run using lme4, version 1.1-7 (Bates, Maechler, Bolker, & Walker, 2013). These models were run twice: once correctly specified given the simulation design, and once with a main effects only model. The model that was correctly specified is meant to serve as the true model, while the main effects only model is meant to serve as a naive implementation that approximates how an actual researcher might approach a dataset for the purposes of exploratory data analysis. Missingness was handled using a single imputation using chained equations, and was calculated using mice, version 2.22 (van Buuren & Groothuis-Oudshoorn, 2011). The missingness model matched the respective statistical model, and thus was correctly specified for the true model and mis-specified for the naive model. Variable importance was only calculated for the main-effects only model, and was simply taken to be the p-value for each respective variable (calculated using a Satterthwaite approximation for the denominator degrees of freedom).

4.2 SIMULATION CONDITIONS

In total, six parameters were of primary interest to be manipulated in the simulation study. The first two, number of levels in the data generation process and nature of the outcome, were fixed to keep this simulation from becoming too unwieldy. The number of levels was fixed to two for simplicity, while the nature of the outcome was fixed to be continuous. Simulation research regarding recursive partitioning methods often do this as a simplifying step, finding that the results of continuous (or categorical) outcomes often generalize to the other condition (e.g., Strobl et al., 2007).

The first parameter of interest is sample size, which can systematically vary at two possible levels of analysis in the context of a two-level design. Common values can also be quite different depending upon the nature of the sample and research questions. For example, a limiting factor for sample size in some education research is the number of units at the first level of analysis (e.g., children in a classroom, or teachers in a school). In the same vein, however, it might be easier to collect units at the first level (e.g., students in a school). To reflect this potential variability, four categories were chosen: 15/20, 15/50, 15/100, 50/100, where the first value corresponds to the level-1 sample size, and the second value corresponds to the level- 2 sample size. Note that level-1 samples within each category were simulated to be unbalanced by sampling from a uniform distribution with a range of 10-20 for the first three categories and 35-65 for the last category. Thus, the total sample size for each condition is approximately the product of the sample size at each level. These values were chosen

to range from a small scale study with a limited sample size at both levels, to a larger scale study with a large sample size for both levels.

The second parameter that was varied was the intraclass correlation coefficient (ICC), which reflects the proportion of variance in the outcome that is attributable to the second level of the analysis. Peugh (2010) reports that education research typically reports ICCs ranging from 0.05 to 0.20. To reflect these common values, three categories were selected for this parameter: 0.0, 0.15, and 0.30. An ICC of 0.0 corresponds to data that are independent, an ICC of 0.15 is meant to represent an average ICC level found in educational research, while an ICC of 0.30 is meant to represent a large ICC level found in educational research.

The third parameter that was varied was the amount of missingness in the covariates. While the mechanism behind missing data can be complex (e.g., Rubin, 1976), this study will only implement data that are missing completely at random. Because of the non-parametric nature of these algorithms and their built-in methods to handle missing data, recent research has found that simulation results for data that are missing completely at random are very similar to results with more complicated missingness mechanisms (Hapfelmeier et al., 2012; Rieger, Hothorn, & Strobl, 2010). Missingness was set at three possible categories for all covariate values: 0%, 10%, and 30%. These values were meant to represent missingness percentages that are small, moderate, and large in the context of educational datasets. Note that missingness in the outcome was not simulated, as there is no straightforward way to handle such missingness in a non-parametric way.

Finally, the level at which the covariates are measured was also manipulated. Given that preliminary simulations showed potential issues with level-2 variables, the variables were simulated to be either level-1 only, or both level-1 and level-2. A condition with only level-2 variables is not needed here, as such an analysis typically aggregates the outcome to the second level, creating a new dataset that is no longer multilevel in nature. Thus, with a fully crossed simulation design, there are 72 ($4 \times 3 \times 3 \times 2$) simulation conditions. See Table 2 for a summary of the main parameters that were manipulated in this study.

Three additional parameters of minor interest were also investigated, but using only one simulation cell each (namely, the cell with a sample size of 15/100, an ICC of 0.15, no missingness, and covariates all at level-1). The first condition was implemented to examine the effect of a balanced hierarchical design. Typically, multilevel datasets are comprised of unbalanced sample sizes nested within clusters, which can negatively impact some models (e.g., repeated measures ANOVAs) more so than others (e.g., growth curve models). This con-

Table 2: *The parameters manipulated in the simulation study.*

Simulation Parameter	Number of Categories	Category Values
Number of levels	1	2
Nature of outcome	1	continuous
Sample size (L1/L2)	4	15/20, 15/50, 15/100, 50/100
Intra-class correlation	3	0, 0.15, 0.30
Missingness (MCAR)	3	0, 10%, 30%
Covariate level	2	level-1 only, both levels

dition was of interest to see how the results may differ between a balanced and an unbalanced design for forest models.

The second condition was implemented to examine to effect of correlations among predictor variables. For simplicity, the relations among the predictor variables were simulated to be independent (more detail regarding data generation can be found in the next section). However, assuming independence among predictors, especially when the number of predictors is high, is often not tenable. One major benefit of forest methods that is often cited is their ability to handle highly correlated predictor variables (Strobl et al., 2008). This condition was of interest to see how the performance of forest methods compared to a multilevel regression model when predictors were no longer independent.

Finally, the third condition was implemented to examine the effect of having a predictor variable that was measured with more measurement error than the others. Theoretical constructs in the social sciences are typically measured with error, and some constructs have more measurement error than others (Novick, 1966). This condition was of interest to determine how the performance of forest models change when more measurement error is introduced.

Results for these special conditions will be displayed separately in the results section below. In total, this resulted in 75 unique conditions, each of which were run 200 times resulting in 15,000 datasets. With six predictive models (CART, CTREE, CART forest, CFOREST, and two multilevel regression models), a total of 90,000 models were run for the simulation phase of this dissertation.

4.3 DATA GENERATION

While the implementation for the data generation varied slightly depending upon which condition was being used, the logic essentially remained the same. In total, 10 variables were simulated of various types. Four variables were continuous, four variables were binary,

and two were Likert type items (one on a 4-point scale and one on a 5-point scale). For all conditions, only two variables were simulated to have a true relationship with the outcome: a continuous variable and a categorical variable (these two variables are hereby referred to as *meaningful variables*). Thus, any detectable relationship between a non-meaningful variable and the outcome will only be due to sampling error (these eight variables are hereby referred to as *meaningless variables*).

The meaningful categorical variable was either a level-1 variable or a level-2 variable depending on the condition, while the meaningful continuous variable was always at level-1. In addition to the main effects for these two variables, a quadratic trend for the meaningful continuous variable as well as an interaction between the meaningful continuous variable and meaningful categorical variable were simulated. The model equations can be seen below in their mixed-effects form, with the model for the "level-1 only" covariate level manipulation presented first followed by the "both levels" manipulation. Note that while the other eight variables are not depicted in these equations for brevity (and because they had no relationship with the outcome), three of these eight were simulated to be either a level-1 variable or a level-2 variable depending on the condition.

Level-1 Only:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{10} X1_{ij} + \gamma_{20} X1_{ij}^2 + \gamma_{30} X2_{ij} + \gamma_{40} X1_{ij} X2_{ij} + \\ & \mu_{0j} + \mu_{1j} X1_{ij} + r_{ij} \end{aligned} \tag{14}$$

Both levels:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{10} X1_{ij} + \gamma_{20} X1_{ij}^2 + \gamma_{01} X2_{j} + \gamma_{11} X1_{ij} X2_{j} + \\ & \mu_{0j} + \mu_{1j} X1_{ij} + r_{ij} \end{aligned} \tag{15}$$

The main effect for X1 (continuous) was fixed to be 0.3, the main effect for X2 (categorical) was fixed to be 0.1, the effect for the quadratic term for X1 was fixed to be 0.2, and the interaction effect between X1 and X2 was fixed to be 0.3. Because all predictors and the outcome were z-transformed, these effect sizes can be interpreted as standardized estimates. The values were selected to represent a moderate deviation from a main-effects only model, with all effects ranging between small and medium according the arbitrary cutoffs of Cohen (1988). Additionally, both the intercept and the slope for X1 were allowed to vary across clusters. For simplicity, these were fixed to be equivalent and have no covariation. Note that because the variance of the outcome and predictors were fixed to be 1, this allowed for an easier calculation of the ICC conditional on the fixed effects estimates.

In previous simulation research that manipulated the ICC (e.g., Maas & Hox, 2005; Moineddin, Matheson, & Glazier, 2007), the ICC was set by ignoring the fixed effects estimates. This results in generating data with a smaller ICC as its typically defined for the intercept-only model than what was desired in the simulation. Because the ICC has such a large impact regarding statistical power, failing to simulate the ICC properly can result in misleading recommendations for applied researchers. See Section A.1 for more information regarding how the ICC was properly calculated in this simulation.

Adjustments were made to the approach outlined above to handle the three special conditions. Recall that the first condition was implemented to examine the effect of the balanced design, and so all clusters were simulated to have the same number of observations. The second condition was implemented to examine the effect of a correlation between two predictors, and so the meaningful continuous variable was simulated to have a correlation of 0.3 with a meaningless continuous variable. Finally, the third condition was implemented to examine the effect of having a meaningful variable measured with more measurement error, and so the meaningful continuous variable was simulated from a standardized uniform distribution rather than a standard normal distribution.

4.4 EVALUATION CRITERIA

Two criteria were chosen to evaluate the statistical performance of each method.

4.4.1 *Proportion variation explained*

This metric is based on mean-squared error, which is a common metric when comparing predictive models with a continuous outcome. It can be calculated by first computing the mean-squared error, dividing this value by the variance of the outcome, and then subtracting this value from one (i.e., $1 - \frac{MSE}{var(y)}$). Note that if the predictions are very bad, it is possible to have negative values. Because the true multilevel model will yield the highest value for this metric, the best-performing methods will be identified by having a similar, albeit smaller, value.

Recall that in random forest methodology, approximately 37% of all observations in a sample will not be chosen in each bootstrap, which is referred to as the OOB sample. Commonly, this metric is used to approximate MSE on a test set without actually needing to split the sample. In a cluster-correlated sample, however, observations within a given cluster are more related to other observations within that cluster compared to observations in other clusters. Because of this, exposing a tree to a given observation within a cluster actually informs the tree about other observations within that same cluster.

This results in trees that are correlated and an OOB error estimate that is overly optimistic (Karpievitch et al., 2009). For this reason, the MSE for proportion variation explained was calculated using a simulated hold out test set.

4.4.2 *Variable importance*

This metric reflects how each method correctly identifies the variables that were actually simulated to have a relationship with the outcome and ranked in the proper order. Given the approach for data generation, variable importance should identify variable 1 as being the most important, followed by variable 2, followed by an eight-way tie for the remaining variables. As mentioned previously, variable importances were calculated via the built-in permutation-based procedure for both forest models and by the corresponding p-values for each variable for the main-effects only multilevel model. Note that because the denominator degrees of freedom are not trivial to compute for multilevel regression models, a Satterthwaite approximation is used in order to estimate the p-values (Goodnight, 1980).

4.5 SIMULATION RESULTS

First, simulation results will be presented for the proportion of variation explained metric. Because a possible confound in these results could be how the methods handled missingness in different ways, results for this metric will only involve conditions with full data. Next, simulation results for variable importance will be presented, aggregated by each condition manipulation. Finally, the results for the three special conditions will be given, first for the proportion of variation metric, followed by the variable importance metric.

4.5.1 *Proportion variation results*

As expected, the naive main-effects only multilevel model performed the worst with respect to predictive ability, followed by the two decision tree models, and then by the two random forest models. A larger discrepancy between the predictive performance of the true model compared to the other models became apparent when the ICC value was high and covariates were included at both levels of the analysis. Sample size also played a large role in prediction, where estimates of the proportion of variation metric became more stable in conditions with larger sample sizes. In fact, negative values for proportion of variation explained are exhibited in the condition with the smallest sample size, largest ICC, and covariates at both levels of the analysis. At least for this simulation study, this indicates that predictive models built on smaller samples of multilevel data are unlikely to yield

predictions that are generalizable to a future sample. Finally, CART yields the poorest performance in the high ICC condition with covariates at both levels for the two smallest sample size conditions. See Figure 9 for the results of the proportion variation explained metric by method, ICC, covariate level, and sample size.

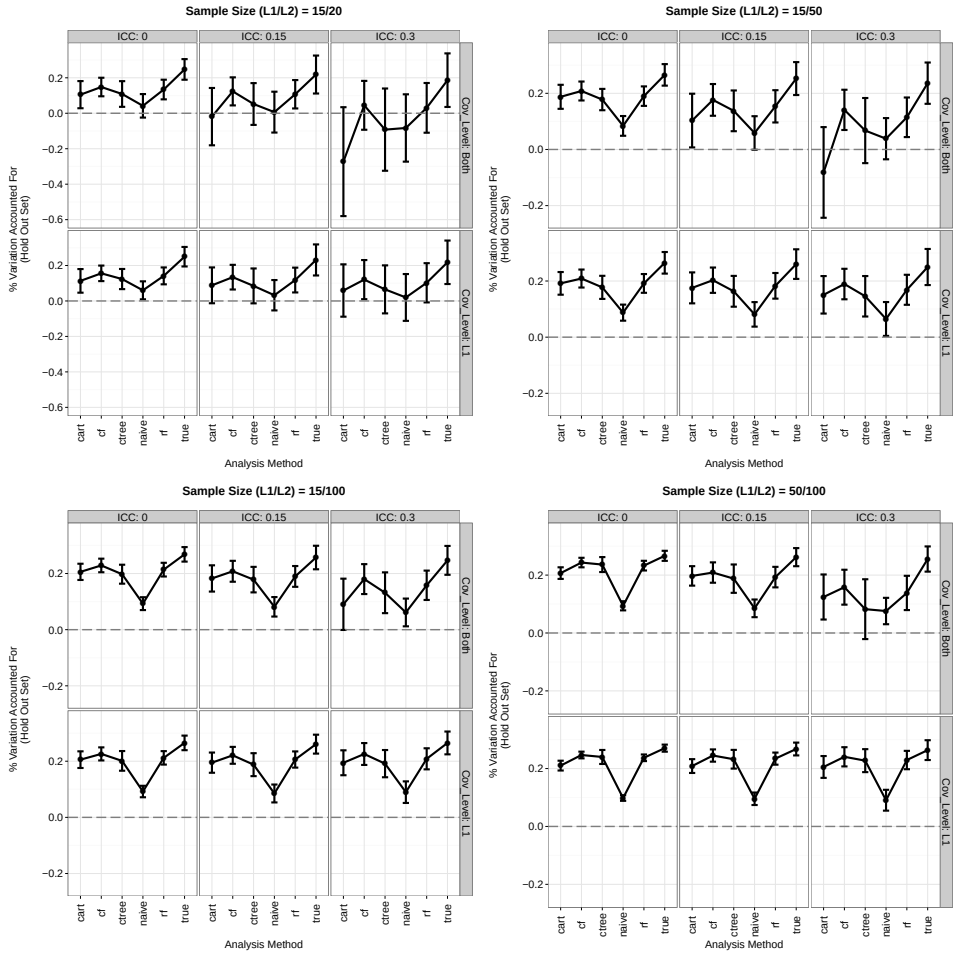


Figure 9: *Proportion variation explained metric by statistical method (x-axis), ICC value (columns), and covariate level (rows). Each plot in the matrix depicts the same thing, but with different sample sizes. Error bars depict the standard deviation of 200 simulations within each condition.*

4.5.2 *Variable importance results*

4.5.2.1 *Sample size*

Recall that for variable importance, the ideal pattern would show variable 1 in first place, variable 2 in second place, and the remaining eight variables in an eight-way tie for third place (and thus earning a rank of 6.5). Despite being miss-specified, the naive model is most indicative of this trend. Both conditional inference trees (ctree) and forests (cf) are close to this desired pattern, which is expected given

they are based on an unbiased algorithm. These conditional inference methods do show some bias, however, for the variables that were sometimes placed at the second level of the analysis depending upon what simulation condition it was (variables 5, 8, and 10). The CART-based methods showed their well-documented bias toward variables that were continuous (variables 1, 3, 4, and 5) or non-binary (9 and 10). Overall, sample size did not seem to play a large role for the variable importance metric like it did with the proportion variation explained metric. See Figure 10 for the results of the variable importance metric by variable, statistical method, and sample size.

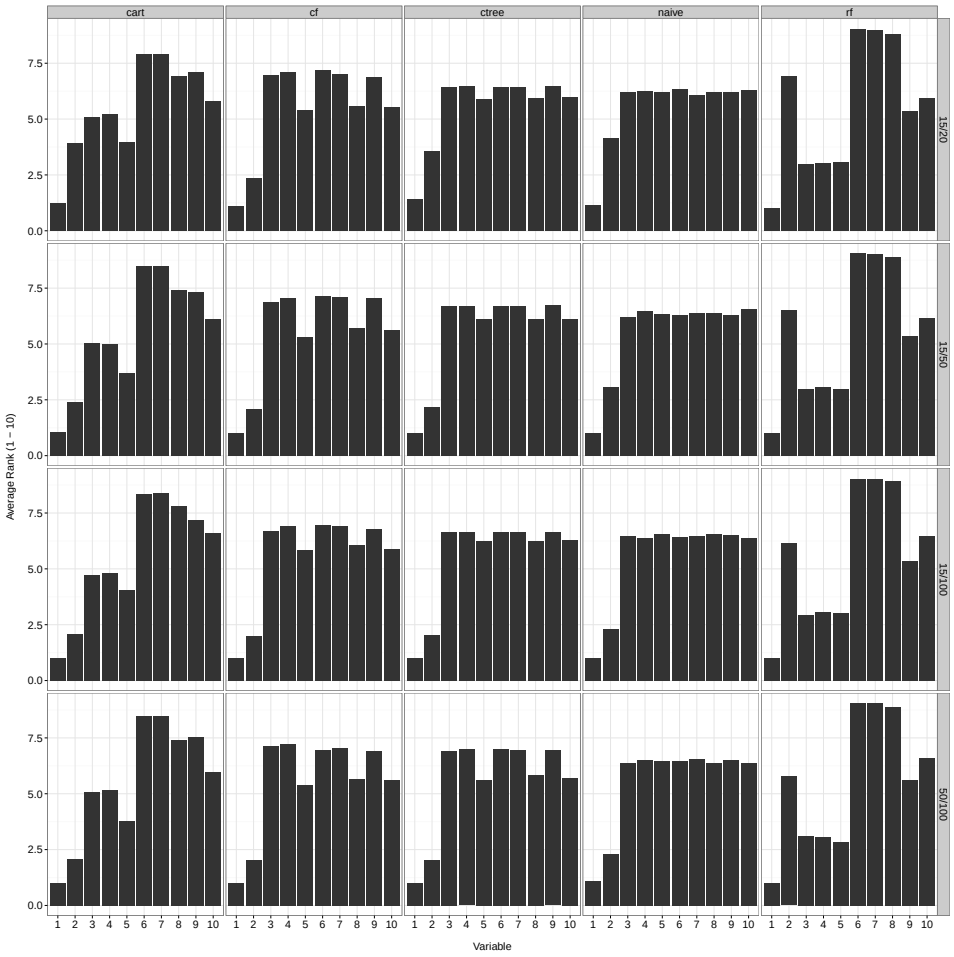


Figure 10: *Variable importance metric by variable (x-axis), statistical method (column) and sample size (row). Ranks were created and averaged across the 200 simulations within each condition.*

4.5.2.2 *Missingness*

Again, despite being miss-specified, the naive model is most indicative of the desired trend. The general pattern among most methods is also consistent across missingness conditions. CART, which was the only method to use surrogate splits, becomes more uniform as the

missingness increases, suggesting the surrogate splits may not work very well with multilevel data and a large percentage of missingness. Conditional inference trees were consistent across missingness conditions, with one very large exception: the power to detect variable 1 as being important completely disappears once missingness is introduced. This suggests that using a CART forest-based imputation procedure is not appropriate to impute data for a conditional inference tree. Both forest methods were consistent across missingness conditions, which suggests that using a CART forest-based imputation procedure for these methods does not result in any problematic behavior for these simulation conditions. See Figure 11 for the results of the variable importance metric by variable, statistical method, and missingness.

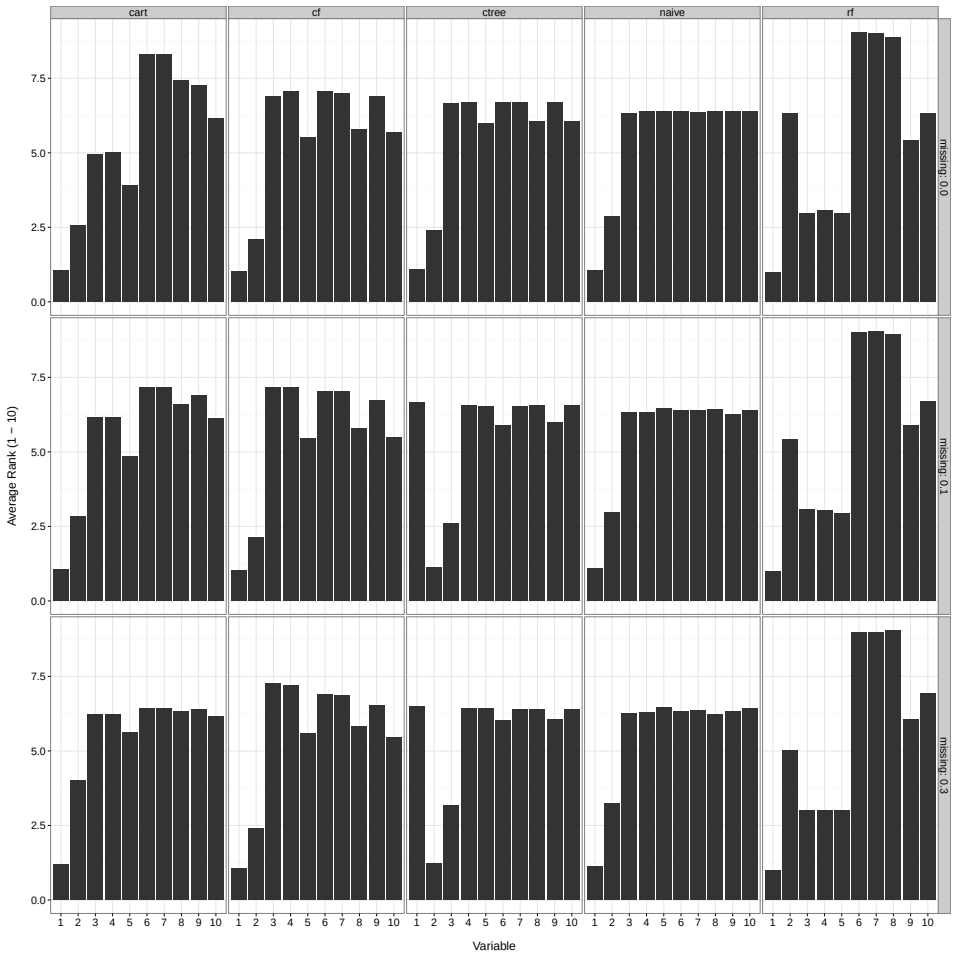


Figure 11: *Variable importance metric by variable (x-axis), statistical method (column) and missingness (row). Ranks were created and averaged across the 200 simulations within each condition.*

4.5.2.3 *ICC and covariate level*

Given the results of the proportion variation explained metric, variable importance results will be discussed for both ICC and covariate level together. Again, despite being miss-specified, the naive model is most indicative of the desired trend. This pattern is also consistent across both ICC and covariate level conditions. All methods show consistent performance across the ICC condition when the covariate level is level-1 only. Additionally, conditional inference methods yield unbiased variable importance measures when the covariate level is level-1 only, regardless of ICC. Bias is introduced to conditional inference methods when covariates are included at both levels of analysis, and this bias is apparent even when the ICC is negligible. Somewhat surprisingly, CART trees and forests show an artificial preference for variables included at the second level of analysis under high ICC conditions, and like conditional inference methods, this bias increases as the ICC increases. See Figure 12 for the results of the variable importance metric by variable, statistical method, ICC, and covariate level.

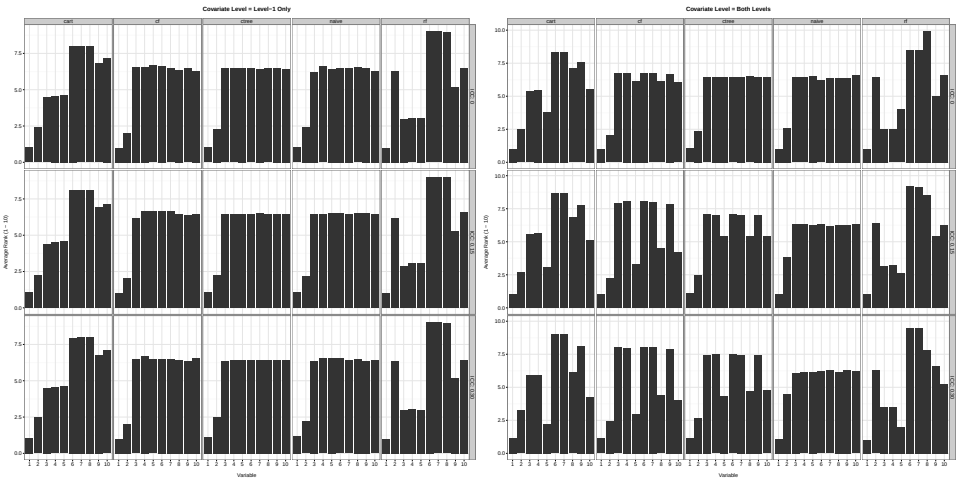


Figure 12: *Variable importance metric by variable (x-axis), statistical method (column), ICC (row), and covariate level (facet). Ranks were created and averaged across the 200 simulations within each condition.*

4.5.3 *Special conditions results*

4.5.3.1 *Proportion variation*

Results regarding the proportion variation explained metric were similar across the condition included in the original simulation and the three special conditions of balanced clusters, a meaningful variable assessed with more measurement error, and a moderate correlation between a meaningful variable and a meaningless variable. Both the balanced clusters condition and the correlation condition yielded no difference compared to the original condition results. The extra measure-

ment error condition yielded a decrease in predictive performance for the true model, as well as a slight decrease for the two forest models. This decrease is to be expected, because when a variable is measured with more error, its relationship with the outcome becomes less reliable and so the effect detected in a training sample will be less generalizable to a test sample. See Figure 13 for the results of the proportion variation explained metric for the three special conditions by method and special condition.

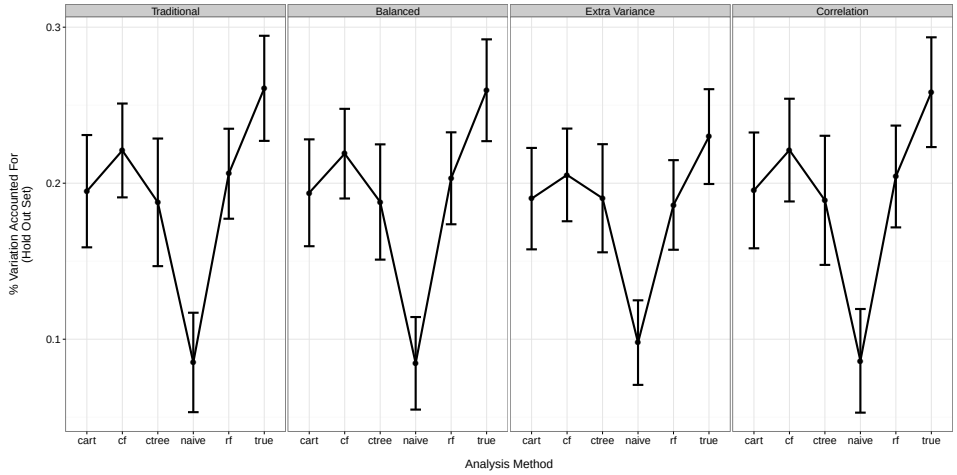


Figure 13: *Proportion variation explained metric by statistical method (x-axis), and special simulation condition (column). ICC value, missingness, covariate level, and sample size (L1/L2) were all fixed to be 0.15, 0% missing, level-1 only, and 15/100, respectively. Error bars depict the standard deviation of 200 simulations within each condition.*

4.5.3.2 *Variable importance*

Much like the proportion variation explained metric, the results regarding the variable importance metric were also similar across the conditions included in the original simulation and the three special conditions. Both the balanced clusters condition and the measurement error condition yielded no discernable difference compared to the original condition results. The correlation condition yielded a slight preference for the meaningless variable (number 3) that was correlated with a meaningful variable (number 1), despite having no simulated relationship with the outcome itself. See Figure 14 for the results of the variable importance metric by variable, special condition, and statistical method.

4.6 DISCUSSION

The goal of this simulation study was to examine the performance of recursive partitioning methods in multilevel data that have similar

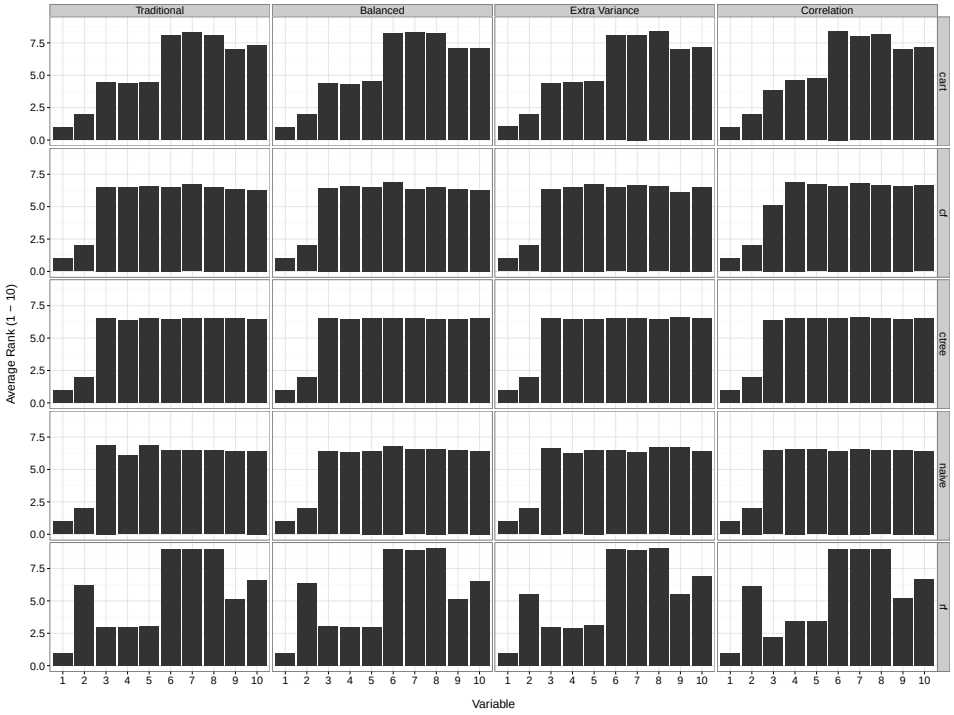


Figure 14: *Variable importance metric by variable (x-axis), statistical method (row), and special simulation condition (column). ICC value, missingness, covariate level, and sample size (L1/L2) were all fixed to be 0.15, 0% missing, level-1 only, and 15/100, respectively.*

characteristics to what is found in traditional educational research. In total, there were three main findings and one minor finding of note. First, with regard to conditional inference methods, results indicated that bias in variable importance is introduced when both level-1 and level-2 variables are included in the splitting procedure, and this bias is exacerbated when the ICC increases. This follows conceptually from traditional regression techniques, where standard error inflation most often occurs due to the presence of level-2 variables incorporated at the first level of the analysis (Luke, 2004). Despite this bias, conditional inference techniques can be expected to yield unbiased variable importance in multilevel data structures where only level-1 variables are of interest, regardless of ICC. As mentioned previously in Chapter 3, research is now beginning to focus on extending the RE-EM algorithm to conditional inference trees (e.g., Fu & Simonoff, 2015). Removing the variation in the outcome due to the cluster level in an iterative way in order to estimate the fixed-effects decision tree structure may help to mitigate this substantial bias when the ICC is high and covariates are included at both levels of the analysis. While this is true for trees, this simulation showed that simply having an ICC of zero is not enough to remove all the bias in variable importance when covariates are at both levels for conditional inference forests. Incorporating the nesting into the permutation framework

through some sort of multilevel permutation procedure or post-hoc correction of standard errors in the conditional inference algorithm could be a potential solution to be investigated for a conditional inference forests.

Second, and most surprising, was the bias found in variable importance for CART-based algorithms. Despite having a simulated relationship with the outcome via a small main effect and a medium interaction effect, variable 2 was not detected as being related to the outcome given it was binary in nature. This indicates that variable importance measures for CART-based methods can be unreliable for small to medium effects, which are effect sizes commonly found in the social sciences. Additionally, these algorithms showed a biased preference for level-2 variables when the ICC was high and covariates were included at both levels, despite theoretical reasons to believe that splits for continuous level-2 variables would be less likely to be selected.

Third, the forest methods showed good predictive performance across all conditions, with the exception of the condition with the highest ICC and covariates included at both levels of the analysis. This is especially true when the sample sizes were large. This finding indicates that these methods could serve as effective analysis tools in multilevel designs when the focus is purely on prediction rather than variable selection. This is especially true when theory might not be able to dictate how variables are related with respect to potential nonlinearities or higher order interactions and the ICC values are negligible.

Finally, a bias in variable importance was detected for variables that were correlated with a meaningful variable, but had no relationship with the outcome. For this simulation, all variables were simulated to be independent for simplicity, which is certainly not realistic for applied datasets. This bias has been well-documented in the literature, and is the exact issue that Strobl et al. (2008) sought to solve with their proposal of conditional variable importance. This conditional importance metric does reduce the bias found in importance measures due to correlated predictor variables, but was not investigated here given that it is much more computationally expensive to calculate and is only available for recursive partitioning methods that use conditional inference. Future research needs to be performed to better understand the impact of complex relationships among predictor variables, and their impact on variable importance metrics in the context of multilevel designs.

Based on the results of this simulation, I propose the following recommendations for applied researchers interested in using these techniques for either exploration or prediction. As a preliminary step, impute missing data if necessary. Then, use the ICC from an intercept-only model in conjunction with the levels at which the study variables

were measured to determine any potential biases that may arise when running a forest model. After that, run both a CART forest and a conditional inference forest to estimate the proportion of variation explained (in the case of a continuous variable) or classification error (in the case of a categorical variable). If the dataset is very large, this can be estimated with a hold out test set. Otherwise, if the dataset is small, k-fold cross-validation at the cluster level can be used, which has been shown to provide approximately unbiased estimates of test error in the presence of dependent data (Rice & Silverman, 1991; Segal, 1992). Finally, if the forest methods perform consistently better, then some evidence exists to warrant the belief that a main-effects only model may be missing some potential nonlinear relationships or interactions that are likely to generalize to a future sample. Otherwise, if the main-effects only model performs similarly or better than the forest models, it is likely that any nonlinear relationships or interactions would either be due to an extremely small effect or statistical noise.

5

APPLICATION PHASE

The purpose of this chapter is to evaluate how the recommendations created using the simulation results from Chapter 4 perform in practice. In total, three datasets were chosen that represent a variety of potential datasets that are typically collected in the field of education. The general approach for all three analyses will follow what was outlined as best practices in the previous chapter. That is, after missing data are handled adequately, the ICC and the levels at which the predictors occur will be considered and compared to the simulation results. Next, the proportion of variation metric for a conditional inference forest and a random forest will be compared to a naive main-effects only model. Finally, based on these results, both variable importance and predicted value plots will be examined for potential nonlinearities and interactions, and subsequent conclusions will be drawn.

5.1 HIGH SCHOOL AND BEYOND SURVEY

High School & Beyond is a nationally representative survey of U.S. public and Catholic high schools conducted by the National Center for Education Statistics (NCES). The data are a subsample of the 1982 survey, with 7,185 students (level-1) from 160 schools (level-2). On average, 45 students from each school were surveyed (Range = 14 - 67). This dataset is publicly available in the nlme package in R (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2015), and is often used for tutorials for two-level multilevel models, most notably in Raudenbush and Bryk (2002). This dataset was chosen for the first application given its simplicity, large sample size, no missing data, and because it is publicly available.

The outcome of interest in this dataset is student math achievement. Possible variables of interest include student race (1 = minority, 0 = other), student sex (1 = female, 0 = male), school size (number of students per school), school sector (1 = Catholic, 0 = public), and the following measures:

SES. A standardized scale of socio-economic status was constructed from variables measuring father's education and occupation, mother's education, family income, and the material possessions of the household. Higher values correspond to higher levels of socio-economic status.

School-level SES. A measure that corresponds to the aggregated SES value of a particular school.

Academic track. A measure that corresponds to the proportion of students in a particular school who are on the highest academic track.

Positive disciplinary climate. A measure indicating the disciplinary climate of the school was based on aggregated student feelings of safety, aggregated student perceptions of fairness and effectiveness of discipline at the school, and the number of discipline incidents among students (e.g., students talking back to teachers, refusal to obey instructions, etc.). Higher values correspond to higher levels of positive disciplinary climate.

High minority status. A measure that corresponds to the proportion of minority students in a particular school, dichotomized such that schools with more than 40% minority enrollment are labeled as having high levels of minority enrollment, while schools with less than 40% minority enrollment are labeled as having low levels.

5.1.1 *Step 1: ICC*

Running an intercept-only model with students nested within schools yields an ICC of 0.18, which indicates that 18% of the variance in math achievement is attributable to the school. Looking at the simulation results, this value is slightly larger than the medium ICC condition of 0.15. Because covariates are included at both levels of the analysis, a potential bias toward level-2 variables could exist in both CART forests and CFOREST.

5.1.2 *Step 2: Estimate proportion of variation explained*

Given the large sample size, the data were split into a training and a testing set rather than performing k-fold cross-validation in order to calculate the proportion of variation in the outcome explained by the statistical model. Refer to Table 5 for the estimates of the proportion of variation metric for the naive model and both forest models. The results indicate that the forest models perform similarly to the naive main-effects only model, with the main-effects only model explaining about 1% more variation in the outcome. This implies that potential nonlinearity or interaction effects are either extremely small or unable to generalize to a future sample.

5.1.3 *Step 3: Examine variable importance and predicted value plots*

Despite the similar performance between methods, both variable importance plots and predicted value plots were examined. Variable importance can be seen in Figure 15. Averaged across the three models, SES and student minority status were found to be the most predictive of student math achievement, while school minority status was found to be the least predictive. This means that despite a bias for level-2

Table 3: *Proportion of variation explained for each method in the High School and Beyond Survey*

Method	Proportion of variation explained (hold out test set)
Naive model	0.20
CART forest	0.19
CFOREST	0.19

variables, both forest methods actually chose level-1 variables as being the most predictive of the outcome. Additionally, both forest models followed similar trends with the exception of the student minority status, which is to be expected given that CART-based methods have a preference toward variables with many potential split points (i.e., not binary). The naive main-effects only model did not have much discriminatory power between variables, which is most likely due to the fact that variable importance was based on p-values, and the sample size for this study was fairly large.

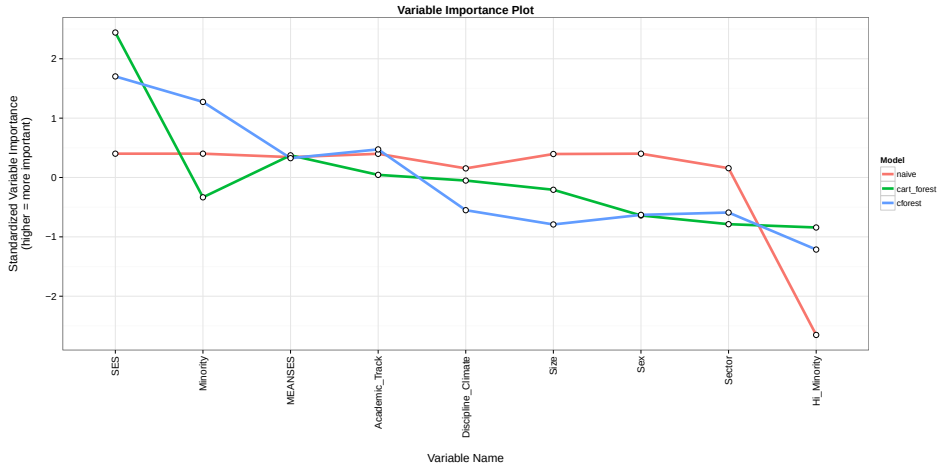


Figure 15: *Variable importance for a naive main-effects only model, a cart forest, and cforest for the High School and Beyond Survey. Note that importance were standardized to allow for easier comparisons. The x-axis is ordered such that the most important variable (averaged across all models) is the furthest left.*

Based on variable importance, predicted value plots were investigated further for SES and student minority status. Plots for these two main effects and their corresponding interaction from the CFOREST model can be seen in Figure 16. The predicted value plot depicts a positive, fairly linear relationship between SES and achievement. Additionally, a negative relationship appears for student minority status, such that minority students tend to have lower achievement levels compared to non-minority students. Finally, there appears to be

an interaction between these two variables, such that non-minority students appear to benefit more from having high SES compared to minority students, who appear to benefit less. However, given the similar performance between the forest models and the naive model, this effect is either very small or unlikely to generalize to a future sample.

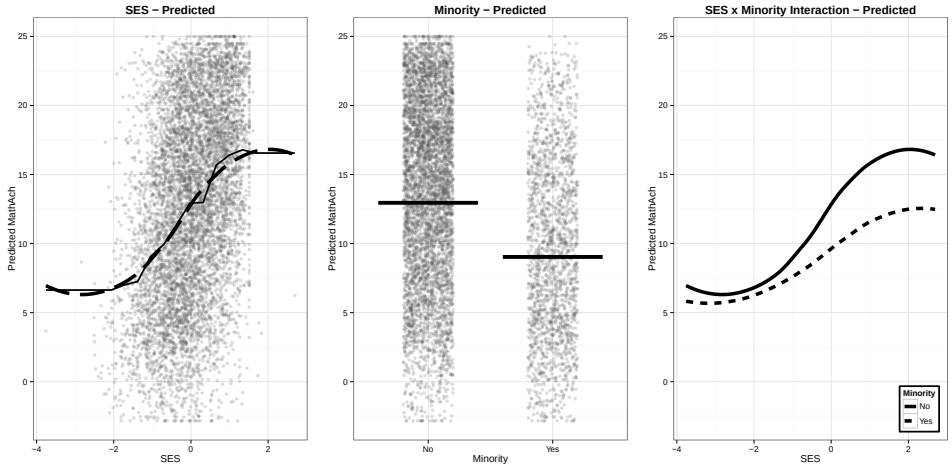


Figure 16: *Predicted value plots for SES and student minority status from a CFOREST model. For SES, the actual prediction is shown with the solid line, while a smoothed loess approximation to this prediction is displayed with the dashed line. The interaction plot just depicts the smoothed approximation for simplicity. Each gray dot seen in both main effects plots represents an observation in the dataset.*

Recall that these plots are simplified versions of partial dependence plots that vary one or two variables while holding the remaining variables constant (see Chapter 2), and that this simplification is made to save on computation time[1]. In this case, the variables were held constant at their median value (in the case of a continuous variable), or most endorsed category (in the case of a categorical value). The values used were from a hypothetical individual who was a female non-minority with an SES of 0, who went to a public school with 1016 students and had 53% of its students on an academic track, a school-level SES of 0.038, and was less than 40% minority. However, it is important to confirm that these results match well with the results from a true partial dependence plot as well as what the raw data might suggest. The predicted value plot using a CFOREST model seen in Figure 16 looks to be a good approximation to both a true partial dependence plot using the same statistical model (seen in Figure 17) and the trends found in the raw data (seen in Figure 18). Additional plots based on a CART forest model which show a similar result can be

1 Indeed this was the case for this dataset, which took over 2 hours to compute the partial dependence data for Figure 17, while the predicted value plot seen in Figure 16 took about 1 second.

found in Section A.2 along with predicted value plots examining the relation between the school-level SES and academic track variables.

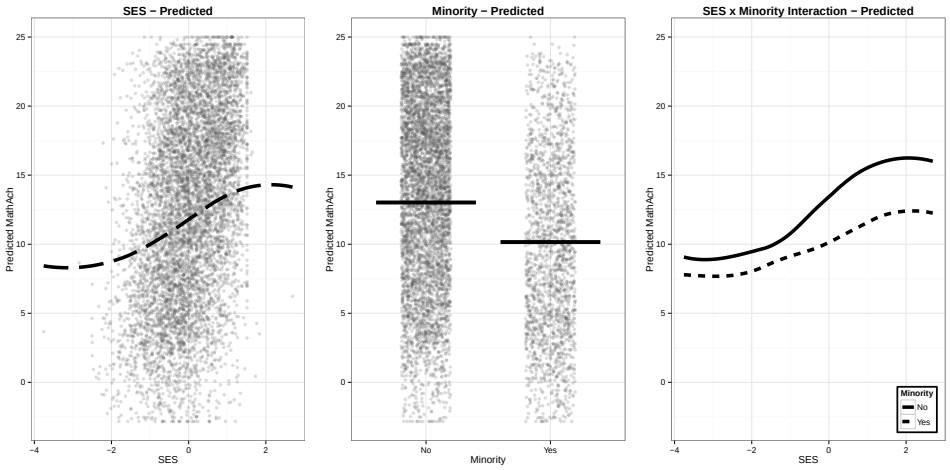


Figure 17: *A true partial dependence plot for SES and student minority status from a CFOREST model. The interaction plot just depicts the smoothed approximation for simplicity. Each gray dot seen in both main effects plots represents an observation in the dataset.*

5.1.4 *Conclusion*

Overall, the main-effects only model performed slightly better (i.e., 1% more variation explained) with respect to predictive performance compared to the two forest models, which performed similarly. This implies that any nonlinearity or interactions that are discovered are most likely very small effects or unlikely to generalize to a future sample. By inspecting variable importance plots, it was determined that both SES and student minority status were potential predictors of interest. Predicted value plots showed that SES appeared to be well approximated by a linear trend. They also showed some evidence for a potential interaction between the SES and student minority status variables, such that non-minority students appear to benefit more from having high SES compared to minority students, who appear to benefit less. Based on the predictive performance comparison, however, this interaction is either extremely small or unlikely to generalize to a future sample. Sure enough, testing such an interaction model on the hold-out test set yields a statistically significant effect ($p = 0.0003$), and the amount of variation explained at level-1 by this interaction is very low ($R^2_{level-1} = 0.003$). In total, there was not much evidence for specifying more complex functional forms beyond a main-effects only model for this dataset. This is not entirely surprising, given that this subset of the High School and Beyond survey is quite popular to use in tutorials, and one of the reasons for this is be-

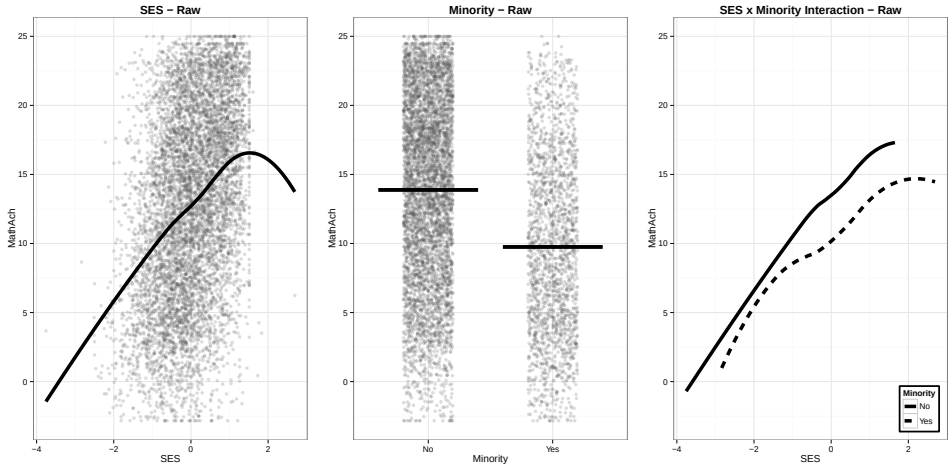


Figure 18: *Raw data for SES and student minority status using a loess smoother. The interaction plot just depicts the smoothed approximation for simplicity. Each gray dot seen in both main effects plots represents an observation in the dataset. Note that the trend predicted low-SES scores is due to the smoother being more affected by an outlying observation than the corresponding forest models.*

cause it is "well-behaved" with respect to being adequately specified by a main-effects only model.

5.2 RESPONSIVE CLASSROOM EFFICACY STUDY

The Responsive Classroom Efficacy Study was a longitudinal, randomized controlled trial examining the impact of a social and emotional learning intervention in upper elementary school grades (Rimm-Kaufman et al., 2014). Schools were randomized into either a treatment group or control group. Third, fourth, and fifth grade teachers in each school either took part in the training program (if in a treatment school), or conducted business as usual (if in a control school). Results examining the efficacy of the Responsive Classroom approach found no statistically significant direct effect of training. However, teacher fidelity of implementation (i.e., the degree to which the training is utilized as intended) was found to be an important mediator of this direct effect, such that treatment assignment predicted teacher fidelity of implementation, which in turn predicted student achievement outcomes.

Various studies have been conducted with the teachers enrolled in this study to further examine the impact of both the Responsive Classroom approach and fidelity of implementation on other outcomes, such as teacher-student interaction quality and the use of standards-based math teaching practices (Abry, Rimm-Kaufman, Larsen, & Brewer, 2013; Ottmar, Rimm-Kaufman, Berry, & Larsen, 2013). These findings suggest that fidelity of implementation is critical to understanding

the contribution of the Responsive Classroom approach on a variety of outcomes. Thus, there is a need for research that examines baseline factors that predict higher or lower levels of fidelity of implementation. In other words, are there signs and signals that can be identified before schools even receive training that may forecast a teacher's uptake of the intervention?

In this current study, recursive partitioning methods are used to explore possible baseline teacher characteristics that might be related to how these teachers implemented their training (if received). The data used for this study are a subsample from the original sample (original N = 350), given that some teachers did not have enough covariate data to perform an imputation. In total, 288 teachers (level-1) from 24 schools (level-2; 13 treatment, 11 control) were included in the sample, with an average sample of 12 teachers per school (Range = 6 - 20). This dataset offers unique methodological challenges with a small sample size, moderate rates of missingness (13% missingness in variables on average), and the majority of predictors being measured at level-1.

The outcome of interest in this dataset is fidelity of implementation of the training program. An important challenge emerged in measuring fidelity of implementation of the Responsive Classroom approach. As the intervention was based on various educational and developmental theories, the training program shares common qualities with what is considered to be typical elementary school practices (Rimm-Kaufman et al., 2014). Because of this, fidelity of the training program was assessed in both intervention and control schools using measures that were free of terminology specific to the training program. Thus, the outcome represents the fidelity of implementation of the training program in intervention schools and teachers' use of practices that resemble Responsive Classroom practices in the control schools.

Fidelity of implementation was created as a factor score via confirmatory factor analysis using three different fidelity measures (sample alpha = .94). The first measure, the Classroom Practices Observation Measure, was an observationally-based measure of teachers' use of Responsive Classroom practices. Research assistants observed and rated teachers and classrooms on 16 items during a one-hour morning observation (sample alpha = .87) and 10 items during a one-hour math lesson (sample alpha = .65). Items were written without training-specific language, so the same assessment was used in both intervention and control schools. Items were rated on a scale ranging from one (not at all characteristic) to three (very characteristic). The second measure, the Classroom Practices Teacher Survey, was a teacher-reported measure of adherence to Responsive Classroom practices consisting of 46 items rated on a 5-point scale, with one indicating not at all characteristic and five indicating extremely characteristics. Items were averaged to create a composite rating (sample

alpha = .91). The last measure, the Classroom Practices Frequency Scale, was also a teacher-reported measure and assessed perceived frequency of the use of Responsive Classroom practices over the course of a school year, consisting of 11 items rated on an 8-point scale with one indicating they almost never used this practice and eight indicating they used this practice more than once a day. Items were averaged to create a composite rating (sample alpha = .88). A factor score was created based on these three measures and used as an outcome in the subsequent analysis.

Three broad categories were identified as potential predictors of interest for this fidelity measure: demographic characteristics, study-specific attributes, and teaching practices and beliefs. The demographic predictors of interest are teacher gender, if the teacher had their Master's degree, years of teaching experience, teacher grade, and treatment assignment.

The study-specific attributes are:

General teacher-student interactions. Teachers rated two items such as "I use hand signals to gain the attention of my class (e.g., raise hand, chime, clapping pattern)" and "I greet each student individually as he/she enters the classroom for the day (e.g., welcome them by name, talk to them)" on a scale ranging from one (not at all character characteristic) to five (extremely characteristic). These items were averaged to create a composite rating, where higher values indicate a higher tendency for these interactions to occur (sample alpha = .29). Note that these two items were among the 46 used to calculate the Classroom Practices Teacher Survey, which was used in the construction of the fidelity factor score.

Use of class meetings. Teachers rated three items such as "In the morning, we have a class meeting where we sit in a circle facing one another" and "I have a predetermined routine in place to structure class meetings (e.g., to determine out greeting, who has an announcement to make, or in what activities we engage)" on a scale ranging from one (not at all characteristic) to five (extremely characteristic). These items were averaged to create a composite rating, where higher values correspond to a higher tendency for these meetings to occur (sample alpha = .71). Note that these three items were also among the 46 used to calculate the Classroom Practices Teacher Survey, which was used in the construction of the fidelity factor score.

The teaching practices and beliefs are:

Teachers' efficacy about their influence in school decision-making. Teachers rated three items such as "How much can you influence the decisions that are made in your school?" and "How freely can you express your views on important school matters?" on a scale ranging from one (not at all) to five (a great deal). These items were averaged to create a composite rating, where higher values correspond to a

stronger belief that they play a role in the school's decisions (sample alpha = .84).

Teacher self-efficacy in relation to instructional strategies. Teachers rated four items such as "To what extent can you craft good questions for your students?" and "How well can you implement a variety of instructional strategies in your classroom?" on a scale ranging from one (not at all) to five (a great deal). These items were averaged to create a composite rating, where higher values correspond to more confidence regarding instructional strategies (sample alpha = .84)

Teacher self-efficacy in relation to classroom management. Teachers rated four items such as "How much can you control disruptive behavior in the classroom?" and "How much can you get students to follow classroom rules?" on a scale ranging from one (not at all) to five (a great deal). These items were averaged to create a composite rating, where higher values correspond to higher levels of classroom management (sample alpha = .92).

Teacher self-efficacy to create a positive school climate. Teachers rated five items such as "How much can you do to make the school a safe place?" and "How much can you do to get students to trust teachers?" on a scale ranging from one (not at all) to five (a great deal). These items were averaged to create a composite rating, where higher values correspond to stronger beliefs in being able to create a positive school climate (sample alpha = .75).

Teachers' attitude toward teaching as a career. Teachers rated eight items such as "I feel I receive the appropriate recognition for my efforts and hard work" and "I feel that I have freedom to make important decisions about my work" on a scale ranging from one (strongly disagree) to four (strongly agree). These items were averaged to create a composite rating, where higher values correspond to more positive feelings toward teaching as a career (sample alpha = .82).

Teachers' perception of principal involvement. Teachers rated 10 items such as "The principal at my school makes clear to the staff his or her expectations for meeting instructional goals" and "The principal at my school sets high teaching standards for teaching" on a scale ranging from one (strongly disagree) to four (strongly agree). These items were averaged to create a composite rating, where higher values correspond to a more positive perception of the principal (sample alpha = .82).

Teacher attitudes toward school programs. Teachers rated five items such as "In my school, you can see real continuity from one program to another" and "In my school, curriculum, instruction, and learning materials are well-coordinated across the different grade levels" on a scale ranging from one (strongly disagree) to four (strongly agree). These items were averaged to create a composite rating, where higher values correspond to more positive attitudes toward school programs (sample alpha = .74).

Teacher attitudes toward Standards of Learning. Teachers rated items about their attitude toward the Virginia state test, the Standards of Learning. The four items included: "In my school, the standards of learning limit the range of instructional practices that I use" and "The standards of learning provide me with instructional targets that inform my instructional focus" on a scale ranging from one (strongly disagree) to four (strongly agree). These items were averaged to create a composite rating, where higher values correspond to more positive attitudes toward standards of learning (sample alpha = .63).

Teacher attitudes toward other teachers. Teachers rated 11 items such as "How many teachers at your school maintain discipline in the entire school, not just their classroom?" and "How many teachers at your school set high standards for themselves?" on a scale ranging from one (almost none) to five (nearly all). These items were averaged to create a composite rating, where higher values correspond to a more positive perception of other teachers at the school (sample alpha = .93).

5.2.1 *Initial missingness step*

The first step to be conducted is to perform data imputation, because this dataset has substantial missingness between teachers who completed baseline measures and those who were still involved with the study one to two years later. Because this is a longitudinal study, attrition is a common occurrence as teachers leave to take new positions, leave the teaching profession altogether, etc. Additionally, the third grade teachers had baseline data and fidelity measured one year apart, while the fourth and fifth grade teachers had baseline data and fidelity measured two years apart. In total, 32% of the sample with baseline data were no longer involved in the study when fidelity was assessed.

Recall that single imputation by the means of a CART forest proximity matrix can only impute values for predictor variables. Thus, imputing data when there is missingness for both the predictors and the outcome requires a two-step process. First, the outcome (32% missing) is imputed via a single imputation using chained equations that takes the cluster level into account via a random intercept of school. Then, with complete data for the outcome, the non-parametric imputation based on the CART forest can impute the predictor values (12% missing on average), resulting in a complete dataset.

5.2.2 *Step 1: ICC*

Running an intercept-only model yields an ICC of 0.65, which indicates that 65% of the variance in fidelity of implementation is attributable to the school. While this value is much higher than what

Table 4: *Proportion of variation explained for each method in the Responsive Classroom Efficacy Study*

Method	Proportion of variation explained (hold out test set)
Naive model	0.39
CART forest	0.35
CFOREST	0.35

is typically encountered in cross-sectional multilevel research (Peugh, 2010), it is not surprising given the context of the research design. Because the fidelity measures were designed to assess teaching practices related to the training program, it is to be expected that teachers who are exposed to training will have higher fidelity. This causes the outcome distribution to be severely bimodal because of an important predictor at the second level (i.e., treatment assignment), thus resulting in a large ICC detected in an intercept-only model. This large ICC should not be problematic with respect to potential variable bias in the forest models, because all but one variable (treatment assignment) was measured at the teacher-level (i.e., level-1).

5.2.3 *Step 2: Estimate proportion of variation explained*

Given the smaller sample size, proportion of variation explained was estimated by using 5-fold cross-validation at the second level of the analysis (i.e., school). Refer to Table 5 for the estimates for the naive model and both forest models. The results indicate that the forest models actually perform worse compared to the naive main-effects only model. This implies that more complex model specifications that involve nonlinearity or interactions either have very small effects or are unlikely to generalize to a future sample.

5.2.4 *Step 3: Examine variable importance and predicted value plots*

Variable importance and predicted value plots were investigated, focusing solely on main effects given that the naive main effects model outperformed the two forest models. Variable importance for each of the 16 predictors can be seen in Figure 19. Averaged across all three models, treatment assignment and the general teacher-student interaction variable consisting of only two items were by far the most predictive, while teachers' attitudes toward school-based programs and whether they have a Master's degree were the least predictive. Both forest models followed similar trends with the exception of grade-level and teacher gender, which is to be expected given that CART-based methods have a preference toward variables that have many po-

tential split points (grade-level and teacher gender had three and two, respectively). The CFOREST model showed no preference among the vast majority of the teacher self-report variables, further solidifying simulation study results that found no variable bias for CFOREST when all variables are measured at level-1.

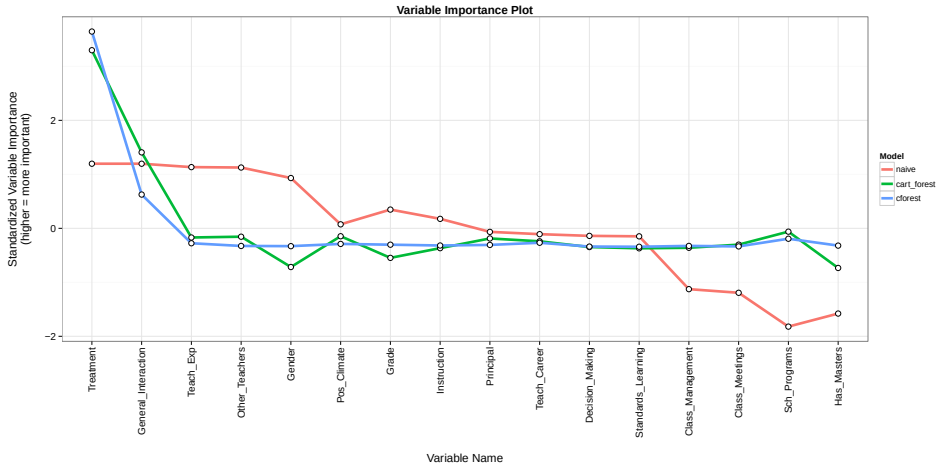


Figure 19: *Variable importance for a naive main-effects only model, a cart forest, and cforest for the Responsive Classroom Efficacy Study. Note that importance were standardized to allow for easier comparisons. The x-axis is ordered such that the most important variable (averaged across all models) is the furthest left.*

Based on variable importance, predicted value plots were investigated further for treatment assignment and general teacher-student interaction. Given the lack of evidence for nonlinearity and interactions, predicted values plots will only show main effects. Plots for these two main effects from a main-effects only multilevel model can be seen in Figure 20, while the same plot using raw data and a loess smoother can be seen in Figure 21. The predicted value plot depicts a large, positive relationship for treatment assignment, such that teachers in intervention schools have higher intervention fidelity than those in control schools. Additionally, the predicted value plot depicts a positive relationship for general teacher-student interaction, such that those teachers who report using hand signals to gain the attention of their class and greeting students individually everyday at baseline tend to have higher levels of implementation fidelity one to two years later. Because this composite had poor reliability indicating these two practices were moderately unrelated in this sample, follow-up correlations were conducted to determine which practice had the stronger relationship with fidelity of implementation. Greeting students at the door ($r = .24$) was found to have a stronger relationship with fidelity compared to using hand signals ($r = .13$).

When compared to the raw data plot in Figure 21, the predicted value plot looks to be a good approximation for treatment, but under-

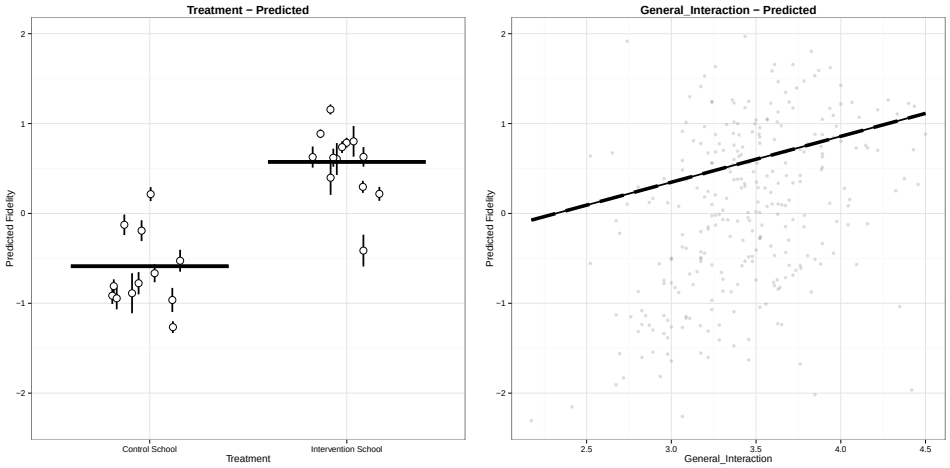


Figure 20: *Predicted value plot for treatment assignment and general teacher-student interaction from a multilevel model. Because treatment was assigned at the school-level, results are aggregated and shown with error bars that represent the standard error of the mean for that particular school. Each gray dot for the general teacher-student interaction plot on the right represents a teacher in the dataset.*

estimates the relation for general teacher-student interaction. Again, recall that predicted value plots are simplified versions of partial dependence plots, where variables are held constant at their median value (in the case of a continuous variable), or most endorsed category (in the case of a categorical variable). Thus, it is important to be wary of the potential bias that can be introduced by this simplification by comparing the result to the true predicted main effect, seen in Figure 22. As evident from the plot, the statistical model is not the source of the bias found in Figure 20; rather, it is the loss of statistical information in the predicted value plotting procedure. This highlights the importance of confirming the result of a predicted value plot with partial dependence plots and raw data plots.

5.2.5 *Conclusion*

Overall, both forest models performed worse than the naive model with respect to predictive accuracy, indicating that there is not much evidence for potential nonlinearity or interactions that are likely to generalize to a future sample. By inspecting variable importance plots, it was determined that both treatment assignment and general teacher-student interactions were potential predictors of interest. Predicted value plots for treatment assignment showed a large, positive relationship with fidelity, such that teachers in intervention schools have higher intervention fidelity than those in control schools. General teacher-student interaction also showed a positive relationship with fidelity, such that those teachers who report using hand signals to

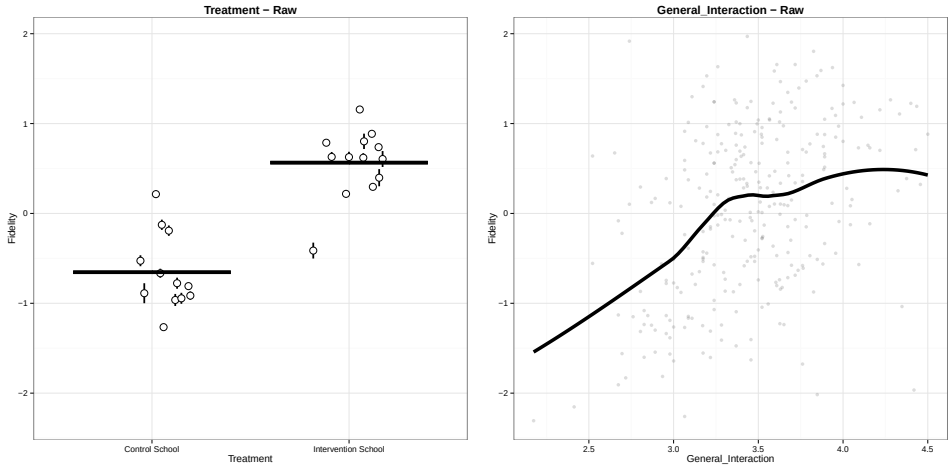


Figure 21: *Raw data plots for treatment assignment and general teacher-student interaction from a loess smoother. Because treatment was assigned at the school-level, results are aggregated and shown with error bars that represent the standard error of the mean for that particular school. Each gray dot for the general teacher-student interaction plot on the right represents a teacher in the dataset.*

gain the attention of their class and greet students individually everyday at baseline tend to have higher levels of implementation fidelity one to two years later.

The first finding indicates the impact of the training program. Teachers randomized to the intervention group showed higher levels of use of the practices. The second finding is somewhat more nuanced and surprising. The Responsive Classroom approach has two hallmark qualities—it is designed to create a caring classroom community and to support teachers' ability to manage behavior effectively. Each of these items in the general teacher-student interaction composite variable corresponds to one aspect of the Responsive Classroom approach. Greeting students as they come into the classroom provides teachers with a way of creating a warm, caring environment. Teachers demonstrate caring toward their students and teachers can observe the students' attitude so that they are able to respond sensitively to the challenges that might emerge during the day (e.g., if the students did not get enough sleep or had a rough time on the bus). Using hand signals is a very effective way of managing the classroom and preventing misbehavior. Hand signals are efficient and proactive and create a channel of communication between teachers and students that help teachers and students shift from one activity to the next and to move efficiently through the managerial tasks in the classroom.

Compared to the first study, both forest models performed much worse in comparison to the main effects only model. There are two main reasons for this. First, treatment assignment played such a large

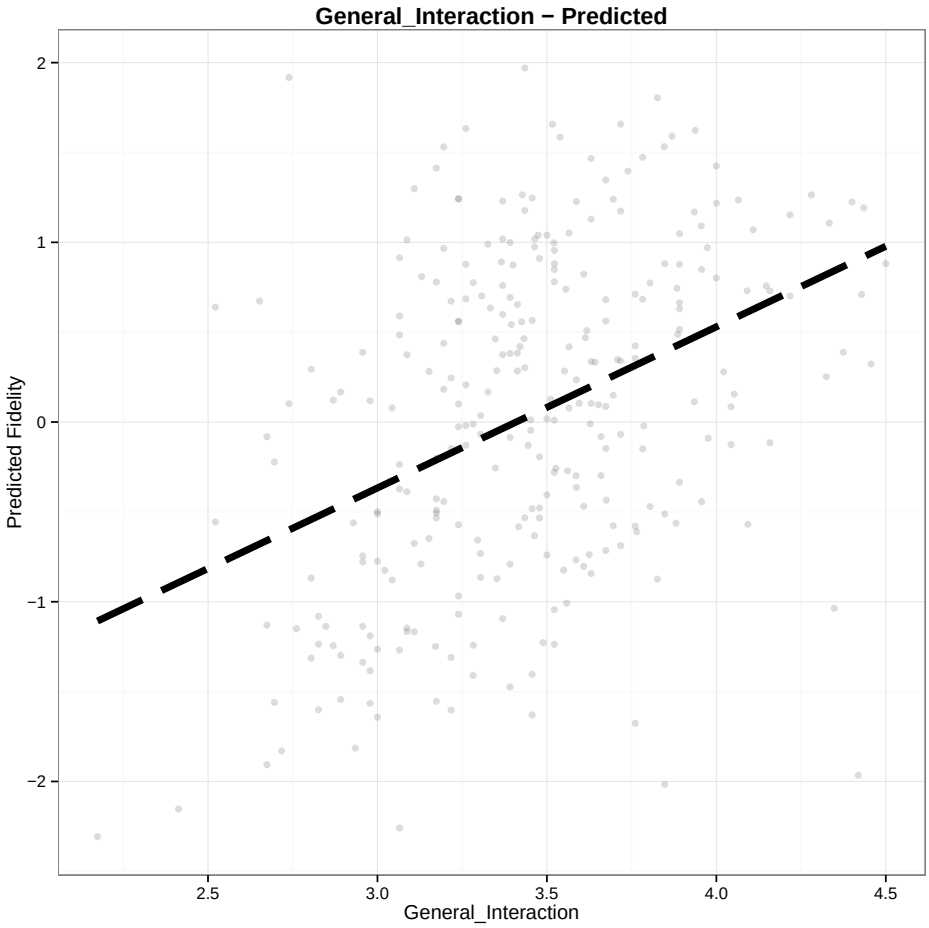


Figure 22: *The true predicted main effects for general teacher-student interaction from a multilevel model. Note that this fit is much more accurate than the corresponding fit from the predicted value plotting procedure seen in Figure 20, indicating that predicted value plots occasionally contain a loss of statistical information and results must be confirmed with true partial dependence plots and raw data plots.*

role in predicting fidelity of implementation, and as a result, there was little systematic variation left to explain. The second reason is the small sample size. As evident in the simulation study, even if nonlinearity and/or interactions exist in the data generating process, forest models have a hard time detecting it in small sample sizes.

5.3 MY TEACHING PARTNER-SECONDARY STUDY

The My Teaching Partner-Secondary Study examined the impact of a Web-mediated intervention focused on improving teacher-student interactions in the classrooms (Allen, Pianta, Gregory, Mikami, & Lun, 2011). Results from a randomized controlled trial indicated that the intervention resulted in student achievement gains in the year follow-

ing completion of the intervention, and this effect was mediated by changes in teacher-student interaction qualities targeted by the training program. In this current study, recursive partitioning methods are used to explore the relation between various student and teacher demographics and psychological constructs measured in the fall and student achievement on a standardized test in the spring. The data used for this study are a subsample from the original study (original N = 2237), using only data collected for students in traditional classrooms (rather than block-scheduled classrooms) in the first year of the project. A total of 1,084 students and 68 different teachers participated in the first year, with classes split across Math (33%), English (32%), History (21%), and Science (14%). On average, there were 15 students per teacher (range = 1 - 28).

The outcome of interest in this dataset is student achievement assessed at the end of the year by the Virginia state standards assessment instrument. Possible variables of interest include prior achievement, student free/reduced price lunch status, student race, student sex, teacher race, teacher sex, teacher age, teacher years of experience, teacher degree, if the class is a high school class, if the class is a math/science class, if the teacher was in the training or control condition, and the following self-report measures:

Teacher-report of school environment. Teachers rated 18 items such as, "School administrators are knowledgeable and competent," and "Teachers help make decisions about things that directly affect them" on a scale ranging from one (never) to six (always). These items were averaged to create a composite rating, where higher values indicate a more positive work environment (sample alpha = .97).

Teacher stress. Teachers rated 12 items such as, "Having little time to prepare" and "Too much paperwork" on a scale ranging from one (no stress/not noticeable) to five (high stress/extremely noticeable). These items were averaged to create a composite rating, where higher values indicate higher levels of work-related stress (sample alpha = .87).

Defiant behavior in class. Students' defiance was assessed using a subscale of the Patterns of Adaptive Learning Scale (Midgley et al., 2000). Students rated four items such as, "I sometimes annoy my teacher during class" and "I sometimes disturb the lesson that is going on in class" on a scale ranging from one (not at all true) to five (very true). These items were averaged to create a composite rating, where higher values indicate higher levels of defiance (sample alpha = .79).

Mastery of achievement goals. Students' mastery of achievement goals was assessed using a subscale of the Patterns of Adaptive Learning Scale (Midgley et al., 2000). Students rated three items such as, "It's important to me that I improve my skills this year" and "One of my goals in class is to learn as much as I can" on a scale ranging from

one (not at all true) to five (very true). These items were averaged to create a composite rating, where higher values indicate higher levels of mastery (sample alpha = .73).

Academic competence. Students' academic competence was assessed using a subscale of the Patterns of Adaptive Learning Scale (Midgley et al., 2000). Students rated three items such as, "I'm certain I can master the skills taught in class this year" and "Even if the work is hard, I can learn it" on a scale ranging from one (not at all true) to five (very true). These items were averaged to create a composite rating, where higher values indicate higher competence (sample alpha = .65).

Peer relationships. Students' ratings of their peer relationships was assessed using a scale developed by Mikami, Boucher, and Humphreys (2005). Students rated four items such as, "How many students in this class do you get along with?" and "How many students in this class respect you and listen to what you have to say?" on a scale ranging from one (I don't get along with anyone in this class/nobody) to five (I get along with everybody in this class/all of them). These items were averaged to create a composite rating, where higher values indicate more positive peer relationships (sample alpha = .67).

Teacher respect. Students rated three items such as, "This teacher interrupts me when I have something to say" and "This teacher never listens to my side" on a scale ranging from one (not at all true) to four (very true). These items were averaged to create a composite rating, where higher values indicate higher levels of perceived respect (sample alpha = .77).

Teacher caring. Students rated three items such as, "This teacher likes me" and "This teacher doesn't seem to enjoy having me in class" on a scale ranging from one (not at all true) to four (very true). These items were averaged to create a composite rating, where higher values indicate higher levels of perceived teacher caring (sample alpha = .66).

Teacher control. Students rated six items such as, "Everyone knows what the classroom rules are" and "The classroom rules are strictly enforced" on a scale ranging from one (strongly disagree) to four (strongly agree). These items were averaged to create a composite rating, where higher values indicate higher levels of rule enforcement (sample alpha = 0.77).

Educational relevance. Students rated four items such as "It's easy to see what we're learning really matters" and "What I learn will probably help me in the future" on a scale ranging from one (not at all true) to five (very true). These items were averaged to create a composite rating, where higher values indicate a higher perception of educational relevance.

Autonomy support. Teachers' ratings of cognitive autonomy support for students was assessed using items created for the MTP-S study (Ruzek et al., under review). Teachers rated six items such as

"Students often get to make decisions about how the class is run" and "Students often feel like they get to help lead the class" on a scale ranging from one (not at all true) to five (very true). These items were averaged to create a composite rating, where higher values indicate higher levels of cognitive autonomy support.

Academic pressure. Students' rated their perceived academic pressure from their teacher using items taken from the Patterns of Adaptive Learning Scale (Midgley et al., 2000). Students rated five items such as "This teacher doesn't let me do just easy work, but makes me think" and "This teacher accepts nothing less than my full effort" on a scale ranging from one (not at all true) to five (very true). These items were averaged to create a composite rating, where higher values indicate higher levels of academic pressure.

This dataset offers unique methodological challenges with a moderate sample size of approximately 1000 students, moderate rates of missingness, and a large number of self-report measures at both levels of the analysis with a wide range in the reliability of these measures.

5.3.1 *Initial missingness step*

The first step to be conducted is to perform data imputation, because this dataset has missingness in both the outcome and predictors. As with the second applied dataset, imputation is performed with a two-step process. First, the outcome (19% missing) is imputed via a single imputation using chained equations that takes the cluster level into account via a random intercept of classroom. Then, with complete data for the outcome, the non-parametric imputation based on the CART forest can impute the predictor values (7% missing on average) resulting in a complete dataset.

5.3.2 *Step 1: ICC*

The hierarchical structure of the dataset in this study actually consists of three-levels: students, teachers, and schools. Because Allen et al. (2011) found no difference between their reported two-level model and a three-level model taking school into account, and the school-level for this study only has an ICC of 0.05, the school-level will be ignored. Running an intercept-only model with two-levels (i.e., children nested within teacher) yields an ICC of 0.47, which indicates that 47% of the variance in student achievement is attributable to the teacher level. Again, this value is much higher than what is typically encountered in cross-sectional multilevel educational research (Peugh, 2010), and is much higher than the large ICC condition in the simulation study. Because the variables of interest were measured at both the student-level and the teacher-level, one can expect bias toward the teacher-level variables to occur.

Table 5: *Proportion of variation explained for each method in the My Teaching Partner-Secondary Study*

Method	Proportion of variation explained (hold out test set)
Naive model	0.16
CART forest	0.24
CFOREST	0.24

5.3.3 *Step 2: Estimate proportion of variation explained*

Given the moderate sample size, proportion of variation explained was estimated by using 5-fold cross-validation at the second level of the analysis (i.e., teacher). Refer to Table 5 for the estimates for the naive model, and both forest models. The results indicate that the forest models explain about 8% more variation in the outcome compared the naive main-effects only model. This implies that more complex model specifications that involve nonlinearity or interactions may be present in the data and are likely to generalize to a future sample.

5.3.4 *Step 3: Examine variable importance and predicted value plots*

Variable importance and predicted value plots were investigated further to identify the source of the additional variation found in the forest models. Variable importance for each of the 25 predictors can be seen in Figure 23. Averaged across all three models, prior achievement of the student by far is the most predictive of student achievement in the following year, while whether the class was a math or science class and the race of the teacher were the least predictive. Both forest models followed similar trends, deviating where expected due to the inherent preference of CART-based methods for variables with many split points. The multilevel model was less discriminating with respect to the prior achievement variable, determining that both prior achievement and free/reduced price lunch status had similar levels of predictive power. Additionally, the forest methods showed a preference for teacher-level variables compared to the multilevel model, further solidifying the findings of the simulation study regarding when the ICC is high and the variables are measured at both levels of the analysis.

Based on variable importance, predicted value plots were investigated further for student prior achievement, free/reduced price lunch status, and teacher-reported school environment. No interactions were detected among these variables, and so the plots will show main effects only in search of potential nonlinearity. Plots from a CART forest

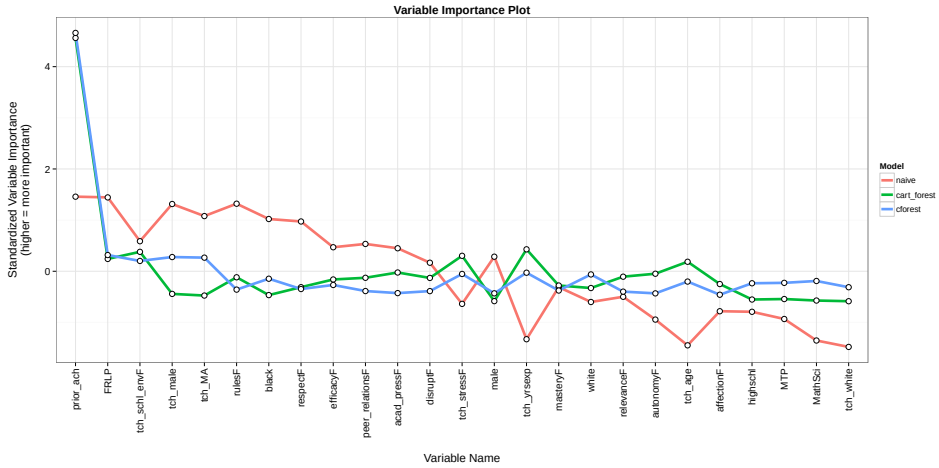


Figure 23: *Variable importance for a naive main-effects only model, a cart forest, and cforest for the My Teaching Partner-Secondary Study. Note that importance were standardized to allow for easier comparisons. The x-axis is ordered such that the most important variable (averaged across all models) is the furthest left.*

model for these three main effects can be seen in Figure 24, while the same plot using raw data and a loess smoother can be seen in Figure 25. The predicted value plot depicts a positive, slightly nonlinear relation for prior achievement, such that students with higher student achievement in the previous year tend to have higher achievement in the subsequent year, and a substantial increase occurs around a prior achievement value of 450. Despite being the second and third most important variables, both free/reduced price lunch status and teacher self-reported rating of the school environment have very little predictive power. A small, negative relation exists for free/reduced price lunch, such that those who are eligible tend to have lower achievement levels. School environment rating has very little if any evidence for predicted relation, though the predicted value plot may be suppressing a curvilinear relation.

When compared to the raw data plot in Figure 25, the predicted value plot looks to be a good approximation for prior achievement, but underestimates the relation for both free/reduced price lunch and teacher-reported ranking of school environment. Again, recall that predicted value plots are simplified versions of partial dependence plots, where variables are held constant at their median value (in the case of a continuous variable), or most endorsed category (in the case of a categorical variable). It is thus important to be wary of the potential bias that can be introduced by comparing the result to a true partial dependence plot, seen in Figure 26. As evident in the partial dependence plot, the predicted value plot serves as a good approximation, but is underestimating a slight curvilinear relation between teacher-reported rating of the school environment, such that

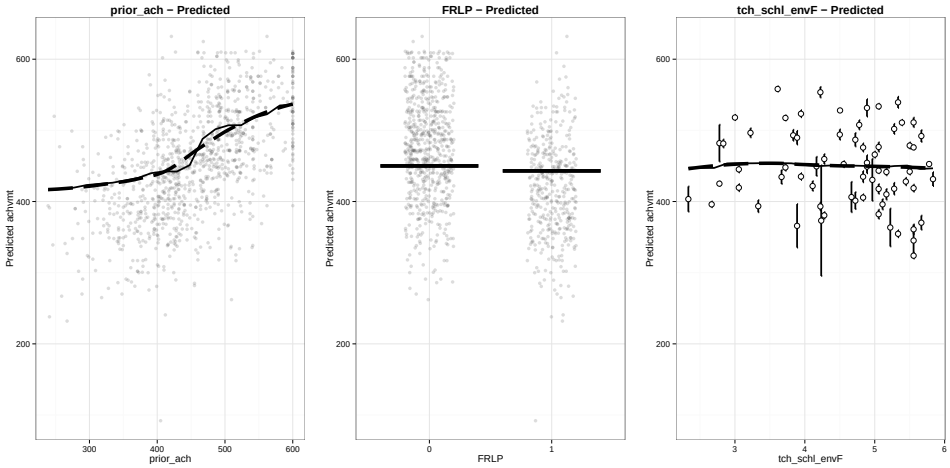


Figure 24: *Predicted value plots for prior achievement, free/reduced price lunch status, and teacher-reported ratings of the school environment from a CART forest model. For prior achievement and school environment, the actual prediction is shown with the solid line, while a smoothed loess approximation to this prediction is displayed with the dashed line. Each gray dot seen in the two leftmost main effects plots represents an observation in the dataset, while results for school environment ratings were aggregated up to the teacher level with error bars that represent the standard error of the mean for that particular teacher.*

a positive relation exists for teachers with low ratings, but then this relation becomes flat for teachers with medium and high ratings.

5.3.5 *Conclusion*

Overall, both forest models outperformed the naive model with respect to predictive accuracy, initially indicating that there is evidence for potential nonlinearity or interactions that are likely to generalize to a future sample. After inspecting variable importance plots, it was determined that prior achievement, student free/reduced price lunch status, and teacher-reported rating of the school environment were potential predictors of interest. Predicted value plots showed a large, positive relation for prior achievement that was also slightly nonlinear. The plots for the other two variables, however, showed almost no relation with the outcome. Additionally, no interactions between these three models were apparent.

This brought up the question: if there was not much evidence for nonlinearity or interactions in the forest models, then why do these models account for 8% more variation in the outcome? This occurs for two reasons. First, prior achievement is such a strong predictor for future achievement. In fact, this variable had the highest standardized variable importance rating for all three datasets examined in this dissertation. Thus, even slight deviations from a linear trend

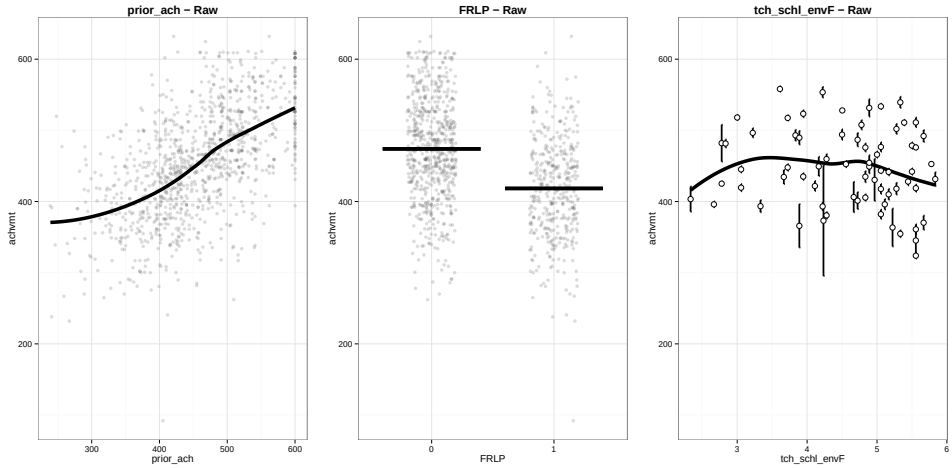


Figure 25: *Raw data plots for prior achievement, free/reduced price lunch status, and teacher-reported ratings of the school environment using a loess smoother. For prior achievement and school environment, the actual prediction is shown with the solid line, while a smoothed loess approximation to this prediction is displayed with the dashed line. Each gray dot seen in the two leftmost main effects plots represents an observation in the dataset, while results for school environment ratings were aggregated up to the teacher level with error bars that represent the standard error of the mean for that particular teacher.*

that are both picked up by a forest model and found to be generalizable to a future sample are likely to account for some of the difference, which is evident in the previous figures. Second, this applied example had a large number of predictors (25), and so this might cause the main effects only multilevel model to overfit, resulting in a poorer cross-validated estimate of variation explained. Sure enough, removing the five least important predictors and refitting the main-effects only multilevel model resulted in a proportion of variation estimate of 22%; now only 2% lower than the forest models. This indicates that while the nonlinearity of prior achievement does improve predictive performance, the effect is likely to be much smaller than the original variation estimates indicate.

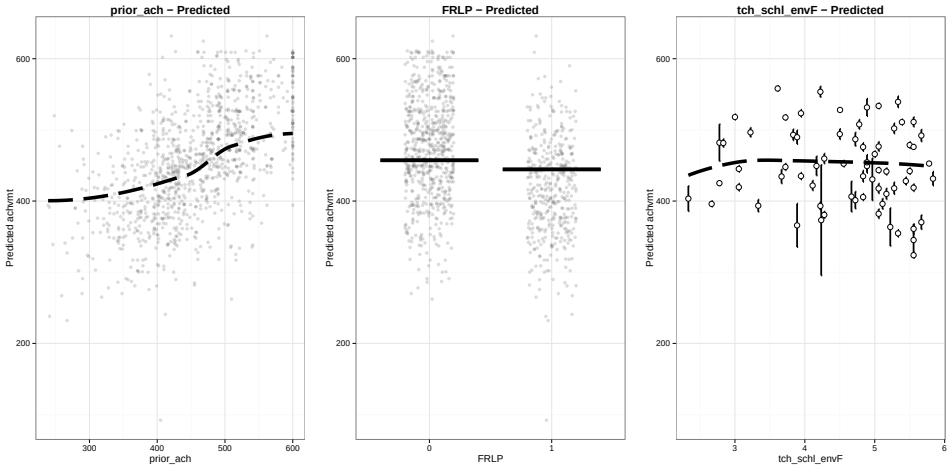


Figure 26: *A true partial dependence plot for prior achievement, free/reduced price lunch status, and teacher-reported ratings of the school environment from a CART forest model. Each gray dot seen in the two leftmost main effects plots represents an observation in the dataset, while results for school environment ratings were aggregated up to the teacher level with error bars that represent the standard error of the mean for that particular teacher.*

6 DISSEMINATION PHASE

Despite the previous findings showing that recursive partitioning techniques, particularly random forests, have the potential to be useful for multilevel exploration in the social sciences, widespread adoption of these techniques will likely be hindered due to a lack of familiarity with (and stigma surrounding) predictive algorithms and "data mining." A dissemination phase was built into this dissertation project in an effort to mitigate this issue, consisting of two main parts: creating an R package and giving a workshop.

6.1 R PACKAGE

An R package, mleda (for **M**ulti-**L**evel **E**xploratory **D**ata **A**nalysis), was created to bundle useful functions assisting in both the exploration of two-level data structures in general, and the application of random forests to such data. Note that decision trees were not included in this package, given that simulation results showed poor prediction performance for tree methods and the application results showed a need to model smoothed functional forms rather than hard, piecewise constants. The code is publicly available for download and installation into R from my GitHub profile (instructions available at `http://github.com/dpmartin42/mleda`). Here, interested users will find helpful introductory R scripts to recursive partitioning methods for both single and multilevel designs, as well as a forum to list potential issues with my code.

This package containes three main functions. The first, `validate_ml`, helps to perform model validation to estimate how well the performance of a given model might generalize to a future sample. This is done in two ways. The first way is with split-half validation. This procedure randomly partitions the dataset into a training set and a test set. The desired model is then built using the training data and performance is calculated on the test dataset. The second way is with k-fold cross-validation. This procedure randomly partitions the dataset in k parts, or folds. The model is then trained on $k-1$ folds, and then tested using the left out kth fold. This step is repeated k times so all folds act as a test dataset, and performance is taken to be the mean performance across the k values. K-fold cross-validation is recommended for smaller datasets (i.e., $N < 1000$), while split-half validation is recommended for larger datasets to reduce the amount of total computation time. For this function, models can either be a random forest (built using the CART or conditional inference algorithm)

or a multilevel model. The models are evaluated using proportion of variation explained (i.e., $1 - \frac{MSE}{var(y)}$) in the case of a continuous outcome, and classification accuracy (i.e., $\frac{(TP+TN)}{(TP+TN+FP+FN)}$, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives) in the case of a categorical outcome.

The second, `importance_ml`, creates the variable importance comparison plots seen in Figure 15, Figure 19, or Figure 23 in Chapter 5. This function can plot and compare variable importance values for an arbitrary number of random forest (built using the CART or conditional inference algorithm) or multilevel models. As mentioned previously, variable importance for forest models are calculated using built-in procedures based on permutation tests to break each variable's relationship with the outcome to determine their predictive strength (refer to Section 2.3.2, for a review). Importances for multilevel models are naively defined as the p-value for the respective parameter. Note that because the denominator degrees of freedom are not trivial to compute for multilevel regression models, a Satterthwaite approximation is used in order to estimate the p-values (Goodnight, 1980). These values are then standardized to allow for easier comparisons, and the resulting plots are ordered such that the most important variable (averaged across all models) is the furthest left on the x-axis.

The third, `plot_ml`, plots the relations between a set of variables and a given outcome, limited to main effects and two-way interactions. These relations can either be based on the raw data itself, or by using predictions calculated from either a random forest (built using the CART or conditional inference algorithm) or multilevel model. In the case of plotting the smoothed relations based on the data itself, a locally weighted scatterplot smoothing procedure is used (i.e., loess), which is a non-parametric way to approximate trends in data by fitting regression models to localized subsets of the data rather than the entire dataset (Cleveland, 1979). Predictions from forest models are typically calculated with partial dependence plots, as outlined in Section 2.3.3. As was previously mentioned, these methods can be quite time consuming to calculate, which is problematic given they are being applied for the purposes of performing efficient exploratory analysis. This function takes advantage of a simplification in the partial dependence procedure originally proposed by Milborrow (2014). Rather than repeating the entire training set for all joint values of the predictor(s) of interest, a new dataset is created that consists of only one observation: the median value for all continuous variables and the most-endorsed level for either categorical or ordinal variables. This new observation is then varied across all the joint values of the predictor(s) of interest. While not a true partial dependence plot, it will typically yield similar results and has the same computation time

regardless of the sample size of the training set. For an example of how much time could be saved by using this simplified procedure, refer to Section 5.1.3. The true partial dependence plot in this example ($N > 7,000$) took over two hours to calculate, while the simplified predicted value plot took less than one second. Both plots are virtually identical.

Finally, two-way interaction plots between two continuous variables were initially created using 3-dimensional graphs, much like Figure 7 in Chapter 2. However, feedback from applied researchers indicated that these plots were difficult to interpret. Additionally, there is no easy way to combine the 3-dimensional object created in R with the other plotting procedures for main effects seen in this dissertation in a well-organized matrix. Because of this, a simplification was made in the plotting procedure for continuous by continuous interactions. This simplification follows the procedures for graphing simple slopes found in Bauer and Curran (2005), by treating one variable as the "moderator" and calculating predictions for values of the moderator at -1SD, 0, and +1SD. While this procedure results in a slight loss of statistical information given the non-parametric nature of the predicted values for two continuous variables in forest models, it creates a plot that is both more familiar and easier to interpret for applied researchers. For readers interested in creating the three-dimensional plots in R, a tutorial can be found on my website (`http://dpmartin42.github.io/posts/partial-dependence-1/`).

6.2 WORKSHOP

In addition to this R package, a small workshop was given on the morning of May 26, 2015 to seven substantive researchers in the Curry School of Education interested in learning more about using recursive partitioning for exploratory data mining. The seven participants who attended consisted of four graduate students, one postdoctoral researcher, and two research scientists, all with varying levels of expertise in both quantitative methods and R.

The teaching plan for the workshop proceeds in three parts. The first part, given in a lecture style, begins with a discussion of exploratory data analysis, focusing on how the workshop attendees use it personally in their own research and perceive it being used by social scientists as a whole. I expect, and found this to be the case in the specific workshop given, that there is general agreement that exploratory data analysis can be extremely useful, but also dangerous due to the fact that most use the same technique for both confirmatory and exploratory research (i.e., null hypothesis significance testing). Using this as a transition, the focus is then shifted on giving an overview of recursive partitioning methods and an outline of best

practices to keep in mind when applying these methods to multilevel data structures.

The second part of the workshop introduces implementing recursive partitioning methods on single-level data for continuous and categorical outcomes in R. This essentially is a guided, line-by-line walkthrough of a well-commented R script, explaining results and graphs while the attendees follow along on their own computers. The examples should be chosen to appeal to the applied researcher to help maintain interest. For this workshop, three datasets were used. The first had a continuous outcome and was the graduation rate dataset used in Chapter 2, where the goal was to predict the graduation rate of colleges based on various college characteristics. The second had a categorical outcome with two categories, and the goal was to predict an individual's acceptance into graduate school based on their undergraduate GPA, undergraduate school ranking, and GRE score. The final dataset had a categorical outcome with three categories, and the goal was to predict the species of a flower using various physical measurements of the flower. In this last example, artificial missingness was induced in order to provide guidelines for how to handle missingness with these methods. To match the simulation study, missingness was only induced in the predictors, and the mechanism underlying the missingness pattern was completely at random.

The last part of the workshop extends recursive partitioning methods to multilevel data in R. Like the previous part, this part is also a guided, line-by-line walkthrough of an R script. Only one dataset was used for this section, which was the High School and Beyond Survey found in Chapter 5. For this example, most of the script was based around the mleda package, and so questions and comments were taken regarding the package in an effort to improve it. No problems were reported installing the package from GitHub, with attendees using a wide range of Mac, PC, and Linux operating systems.

After the conclusion of the workshop, open-ended feedback was collected via email so attendees could comment on what they found helpful, what they found confusing, and how these methods might be useful in their own work. The responses from all seven participants can be seen below.

> I had the opportunity to attend the first portion of the presentation, and I was excited to learn about these innovative and useful data analytic methods! You very effectively relayed new and complex information while consistently checking to see if your audience had questions/comments. I also felt you nicely balanced text with visuals (figures/tables) to help convey the meaning of the information and analyses. Overall, great work!

I enjoyed the workshop a lot. I have little experience with R beyond the quant sequence but the way the information was presented was not only logical but also easier to understand than what I've gotten in the past. Having the script (rather than having to recreate it myself) was very helpful– I could follow along and understand the process but not get held back by code problems. The difficult piece was just imagining its application to my own work, but that's something I can sit down with on my own time and is also probably due to my being newer to dealing with big data. I looked back at the R script later and it still seemed clear, and it looks like I could apply it to my own data without huge issues.

Thanks to you for offering the workshop! The CART forests approach is something that we might use in selecting the optimal set of covariates in the estimation of propensity scores or in our exploratory analyses of variables associated with implementation fidelity, and it was presented in a very accessible way. I got a little behind when I was trying to load the various R programs during the workshop, but everything was so clearly labelled in the code that it was really easy for me to go back and use it after the workshop was over. I think I would have benefited from doing some reading prior to the workshop - maybe the Breiman article - but I also know you can't reliably expect workshop participants to do prework.

I enjoyed the hands-on nature of the workshop. It allowed me to follow-along more easily as we worked through the examples. I appreciated how responsive you were to questions and how you were quick to help us troubleshoot if anything went wrong. I think one of the more confusing issues around this work relates to handling missing data, and in particular when you need to do multiple imputation (i.e., when you are missing on the dependent variable?) and how best to do that. Forest methods are something I could see using in my own work when I have a data source with a lot of different measures. Specifically, I would use it when I have exhausted theoretically-informed analyses, but still want to know if there are other predictors that have a direct association with my outcome of interest or interact with other theoretically expected predictors.

> Overall I found the workshop to be very useful. I liked that most of the time was spent hand-on, and that you had the code ready for us. I also found your R package to be very useful, not only for random forests, but the plotting functions will also be useful for other multilevel applications. There was a lot of information during the workshop, and it will help to have the code to go back to. You might consider incorporating your code into tutorials that walk people through the steps again if you plan to continue giving this workshop.

> I really enjoyed Dan's workshop. His R package was super easy to use, and really intuitive. I don't know much about random forests, but Dan explained it so well that I think I can already begin to use his package in R. The workshop was particularly relevant for me because I do a lot of education research, so I'm excited to use his package to explore relationships I haven't yet looked at in the datasets that I work on.

> I found the workshop to be very helpful! It was great exposure to a topic that I was previously unfamiliar with. You did a nice job thoroughly explaining the recursive partitioning method in a very accessible way. Although I don't have immediate plans to use the method, it is definitely something I will keep in mind going forward. I imagine that it will be especially relevant for exploratory analyses. As you know, I am not an R user, but I appreciated you for providing the syntax and walking us through the process of running the models step-by-step. The R package you created is very impressive and will prove helpful when I use this technique in the future.

Overall, this feedback indicates that the workshop was well-received, and the important concepts were understood. More importantly, many participants explicitly expressed confidence that they could run these models in the future and intended on doing so. The only critique was with regard to best practices for handling missingness for multilevel data. While I discussed this, I had to do so briefly to fit everything in the allotted time frame of three hours. As will be discussed in the next chapter, best practices to handle missingness in multilevel designs, especially with respect to recursive partitioning methods, remains an open area of research.

With respect to the R package, mleda, all attendees reported positive feedback with respect to the easy-to-use code and quality of the corresponding plots. One participant even mentioned that this R package is also great option for traditional exploratory data analysis in general, without even needing to use forest models. One of the goals I kept in mind when building this package was to create a product that had a low barrier of entry, so researchers did not even need to know recursive partitioning methods to be able to find it useful. Based on this initial feedback, it looks like this goal has been accomplished. For interested readers, all workshop materials including slides and R scripts are publicly available on GitHub (at `http://github.com/dpmartin42/mleda_workshop`). The multilevel tutorial R script was also incorporated as supplemental information for a manuscript submitted for widespread dissemination in a special issue of *Psychological Methods* on Big Data in the social sciences.

7

GENERAL DISCUSSION

The main goal of this dissertation was to evaluate the feasibility of using recursive partitioning methods as an efficient means to perform exploratory data analysis for multilevel data structures commonly found in the social sciences, and if so, help applied researchers incorporate these methods into their own statistical toolbox. This central question was investigated with three distinct phases: a simulation phase, an application phase, and a dissemination phase. After briefly restating the main findings from each section, limitations and future directions of this work will be discussed.

7.1 SIMULATION PHASE

The goal of the simulation phase was to examine when recursive partitioning methods perform as expected and when they begin to break down in the presence of multilevel data. This dissertation focused on decision trees and forests that used one of two popular recursive partitioning algorithms: CART and conditional inference trees. Both CART and conditional inference methods showed decreased performance in predictive accuracy and the identification of relevant variables when the ICC was moderate to large and predictors were measured at both levels of the analysis. In particular, both methods had a biased preference for level-2 variables, despite these variables having no simulated relationship with the outcome. While this is to be expected with conditional inference methods that utilize a permutation test framework built on independence assumptions, this finding is unexpected for CART methods. If all variables are measured at the first level of analysis, however, both CART and conditional inference methods perform as expected, regardless of ICC values.

7.2 APPLICATION PHASE

The main goal of the application phase was to examine how the findings from the simulation phase would generalize to real data. Results from three separate applications indicated that forest methods did not massively outperform a main-effects only model in any application, but it did aid in the potential identification of small effects that deviated from linearity. In the first application, a very small interaction effect was discovered between the SES and student minority status variables, such that non-minority students appeared to benefit more from having high SES compared to minority students, who

appeared to benefit less. The second dataset had no evidence for anything more complex than a main effect, which is not surprising given the fact that the dataset was small ($N < 200$), making it more difficult to identify more complex model specifications that were likely to generalize to a future sample. The third dataset found a small nonlinearity in student prior achievement predicting future achievement. While the deviation from nonlinearity was small, it contributed to the forest models outperforming the main effects only model due to the fact that the impact of prior achievement on future achievement was large. However, because prior achievement explained so much variation in future achievement, it left very little systematic variation to be explained by other measures.

7.3 DISSEMINATION PHASE

The main goal of the dissemination phase was to examine what applied researchers think about data mining methods in general and recursive partitioning methods specifically. Overall, the feedback from the small subset of researchers I have discussed this research with has been positive with regard to adopting these methods in their own research. An R package and tutorial scripts have been created and are publicly available to help further reduce the barrier of entry for researchers interested in learning more about these methods. Given the relative success of the workshop, I am more than confident that any researcher interested in these methods can learn the basics with only a few hours of reading, regardless of statistical expertise.

7.4 LIMITATIONS AND FUTURE DIRECTIONS

Like with any simulation research, the findings from Chapter 4 are limited in generalizability to the parameter manipulations in this study. While an attempt was made to make conditions as realistic as possible, various parameters of interest were left out to keep the simulation from becoming too unwieldy. For example, parameter values were selected in order to mimic what is traditionally found in cross-sectional education research. Other areas in the social sciences or in longitudinal research, for example, would require different parameter values. A simulation based on longitudinal research would need to have a much larger ICC than what was simulated here, as more variation is typically due to the second level (i.e., between persons) in such studies. Applying these methods to longitudinal studies would make for an interesting future direction, as nonlinear growth patterns are often of substantive interest to developmental researchers (Grimm, Ram, & Hamagami, 2011), and are easy to detect in an exploratory way with recursive partitioning methods.

An additional limitation is the way missingness was simulated. In this study, missingness was restricted to be both missing completely at random and missing only in the predictors and not the outcome. Despite the fact that previous research has found both surrogate splits and imputation by proximity matrix perform admirably across various missingness mechanisms (Breiman, 2003; Hapfelmeier et al., 2012), it is important to replicate that finding for the multilevel contexts investigated here. As for limiting the missingness to only impact the predictors, this was done because the built-in methods to handle missingness for recursive partitioning are limited to missingness in the predictors. This is also why missingness in Chapter 5 was handled via a two step process, using a single, main-effects only multilevel imputation using chained equations for the outcome, followed by a non-parametric imputation based on a proximity matrix for the predictors. Surely this is not an ideal solution, because it combines non-parametric imputation with a potentially miss-specified parametric imputation model. This approach likely suppresses many non-parametric relations between predictors and the outcome, resulting in a higher rate of false negatives. Unfortunately, however, to the author's best knowledge there is no non-parametric way to handle missingness in both the predictors and the outcome at this time. Future research in this direction would be highly valuable; handling missingness both quickly and correctly for exploring either single-level or multilevel data structures using recursive partitioning methods remains an open area of inquiry.

Finally, the last limitation to be discussed is the datasets used in Chapter 5, all of which had potential reasons as to why there was only minimal evidence supporting a departure from linearity. The first sample, the High School and Beyond Survey, is a publicly available subsample from a larger study that is often used in tutorials due to the fact that it is so "well-behaved" with respect to being adequately specified by a main effects only model. The second sample, the Responsive Classroom Efficacy Study, has a small sample size with only two variables accounting for the majority of systematic variation in the outcome. Lastly, the third sample, the My Teaching Partner-Secondary Study, used future achievement as the outcome with prior achievement as a potential predictor. When it comes to predicting achievement, prior achievement will always be the best. In fact, research has found that growth mixture models based solely on trajectories of prior student achievement yields the best performance for predicting future student achievement and dropout in early warning systems. (Knowles, in press).

Note that recursive partitioning was specifically chosen as the focus of this dissertation given that it is nonparametric in nature, relatively easy and intuitive to understand, and their ensemble method counterparts (i.e., forests) consistently show good performance in pre-

dictive tasks when compared to other popular methods (Caruana & Niculescu-Mizil, 2006). However, these methods are certainly not a "silver bullet" for exploratory data analysis or predictive modeling in general. As evidenced by the three applications shown here, linearity is often a valid assumption to make in many statistical applications. Regularized regression methods that include a penalty term in its estimation procedure, such as the L2 norm (i.e., the sum of the squared, standardized parameter estimates) in the case of ridge regression or the L1 norm (i.e., the sum of the absolute value of the standardized parameter estimates) in the case of the LASSO, can be attractive options to reduce overfitting in linear models at the small cost of increased bias (Hastie et al., 2009). While some research has extended these methods into a mixed-effects framework (e.g., Eliot, Ferguson, Reilly, & Foulkes, 2011; Schelldorfer, Bühlmann, & van de Geer, 2011), such a model is quite complex and unlikely to be adopted by applied researchers for the sole purpose of exploratory data analysis. However, if the focus is on prediction, such multilevel extensions may not be necessary. Because the hierarchical nature of the dataset affects the standard errors rather than the regression weights, predictions are likely to remain unchanged regardless of whether the models account for the nested structure of the data or not. Future research should examine this idea more carefully, and compare the predictive performance of forests, regularized regression methods, and a naive main effects only model at various deviations from linearity to see which methods perform the best.

7.5 CONCLUSION

This dissertation examined new ways to conduct exploratory research on multilevel datasets in the social sciences using recursive partitioning. Using both simulation and applied datasets, I have shown that these methods provide a cost-effective way to revisit old datasets to discover something new in a data-driven way, without the hassle of relying theoretical justification (which may or may not exist) to decide what model specifications to try. Taking such an approach for purely exploratory research has the potential to help researchers make new discoveries and inform future research in an efficient manner, all while controlling for potential false positives through the use of validation techniques.

A

APPENDIX

A.1 PROPER CALCULATION OF THE ICC FOR A MULTILEVEL SIMULATION

When the ICC is a parameter of interest in a multilevel simulation study, it is important generate the data to yield the desired ICC as it is typically defined. That is, the variation in the outcome due to the clustering variable in an intercept-only model. Previous multilevel simulation research (e.g., Maas & Hox, 2005; Moineddin et al., 2007) set the ICC by ignoring the fixed effects estimates. In other words, the ICC is a value not defined as above, but rather one which corresponds to the variation in the outcome due to the clustering variable in a statistical model with all the predictors included. This results in a smaller ICC value for the intercept-only model, which can be misleading when it comes to practical recommendations.

Below are guidelines to follow in order to properly calculate the ICC for a multilevel simulation study when the ICC is defined for the intercept-only model. For brevity, these instructions will apply to the level-1 only model used in this dissertation, but the logic can easily be extended to apply to any multilevel model. First, recall the level-1 only multilevel model that was used. This model includes an intercept, a main effect for X1, a quadratic term for X1, a main effect for X2, and an interaction between X1 and X2. Both X1 and X2 were z-transformed in order to yield standardized effect size estimates. Additionally, both the intercept and the slope for X1 were allowed to vary across clusters. For simplicity, these were fixed to be equivalent and have no covariation. The mixed-effects model equation for the level-1 only covariate level condition can be seen below:

Level-1 Only:

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{10}X1_{ij} + \gamma_{20}X1_{ij}^2 + \gamma_{30}X2_{ij} + \gamma_{40}X1_{ij}X2_{ij} + \\ & \mu_{0j} + \mu_{1j}X1_{ij} + r_{ij} \end{aligned} \tag{16}$$

Because X1 and X2 were standardized, the outcome, Y, also needs to be standardized in order to allow the regression weights to be interpreted as standardized effects. Thus, the total variance of both sides of the equation are equal to one. Because the ICC is defined as the variance in the outcome due to the clustering variable over the total variance of the outcome, and the total variance of the outcome is

1 by definition, then the variance in the outcome due to the clustering variable can be fixed to the desired ICC value itself. Now, the only remaining unknown is the variance of the residual term. That is,

$$1 = Var(\gamma_{00} + \gamma_{10}X1_{ij} + \gamma_{20}X1_{ij}^2 + \gamma_{30}X2_{ij} + \gamma_{40}X1_{ij}X2_{ij} + \mu_{0j} + \mu_{1j}X1_{ij} + r_{ij}) \tag{17}$$

Because all predictors were simulated to be uncorrelated,

$$Var(r_{ij}) = 1 - Var(\gamma_{10}X1_{ij}) - Var(\gamma_{20}X1_{ij}^2) - Var(\gamma_{30}X2_{ij}) - Var(\gamma_{40}X1_{ij}X2_{ij}) - Var(\mu_{0j}) - Var(\mu_{1j}X1_{ij}) \tag{18}$$

Now, two facts can be used to further reduce the equation above. First, the regression weights for all slopes (i.e., all gammas except γ_{00}) are scalar values, and so the variance rule of $Var(aX) = a^2Var(X)$ can be applied. Second, a squared standard normal distribution is chi-square distributed with 1 df, and so it has a variance of two. So, knowing that the variance of both X1 and X2 were simulated to be one (i.e., standardized), the equation above becomes

$$Var(r_{ij}) = 1 - \gamma_{10}^2 - 2\gamma_{20}^2 - \gamma_{30}^2 - \gamma_{40}^2 - 2ICC \tag{19}$$

Assuming a desired ICC of 0.3 along with an effect of 0.3 for X1, 0.2 for $X1^2$, 0.1 for X2, and 0.3 for X1X2 with the above level-1 only model, the between-level variance would be set to the ICC value of 0.3 and the within-level variance would be calculated as 0.13. As long as the intercept is fixed to $-\gamma_{20}$, using the above values for the ICC would result in both the parameter values being simulated correctly as well as the ICC for the intercept-only model.

A.2 EXTRA PLOTS FOR THE HIGH SCHOOL AND BEYOND SURVEY

First, both the CART forest and multilevel model analogous to Figure 16 are shown. Both the CART forest (seen in Figure 27) and multilevel model (seen in Figure 28) are similar to what was displayed in Figure 16 with respect to the main effects. Because of the lack of evidence for an interaction, it appears that the interaction plot (or lack thereof) for the multilevel model provides a better model for prediction.

Next, both school-level SES and proportion of students on the academic track were investigated in a similar fashion, because they were the third and fourth most important variables, respectively. Results for a CART forest can be seen in Figure 29, results for a CFOREST can

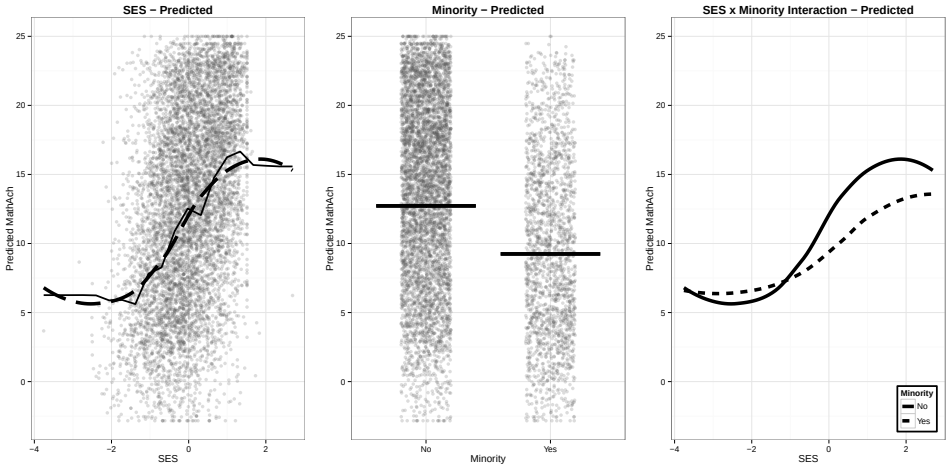


Figure 27: *Predicted value plots for SES and student minority status from a CART forest model. For SES, the actual prediction is shown with the solid line, while a smoothed loess approximation to this prediction is displayed with the dashed line. The interaction plot just depicts the smoothed approximation for simplicity. Each gray dot seen in both main effects plots represents an observation in the dataset.*

be seen in Figure 30, and results for a multilevel model can be seen in Figure 31. All three models show a small, positive relationship with math achievement that seems to be well-approximated by a linear trend, with no evidence for an interaction. However, these trends do not seem to be as strong as the raw data with a loess smoother might suggest (seen in Figure 32). There are two reasons for this. First, there are very few observations for the lower values of both predictors, and the loess is more susceptible to being affected by outlying observations with high leverage compared to a forest or multilevel model. Second, it is because these plots are predicted value plots, and thus sacrifice a loss of statistical information for computational efficiency.

This can be confirmed without the hassle of computing partial dependence plots for the forest models by comparing the predicted value plots of the multilevel model to what would be predicted using the training set. These results can be seen in Figure 33, which show the true predicted relationship of both main effects. As evident from the plot, the statistical model is not the source of the bias shown in the previous plots, it is the loss of statistical information in the predicted value plots. This highlights the importance of confirming the result of predicted value plots with true partial dependence plots and raw data plots.

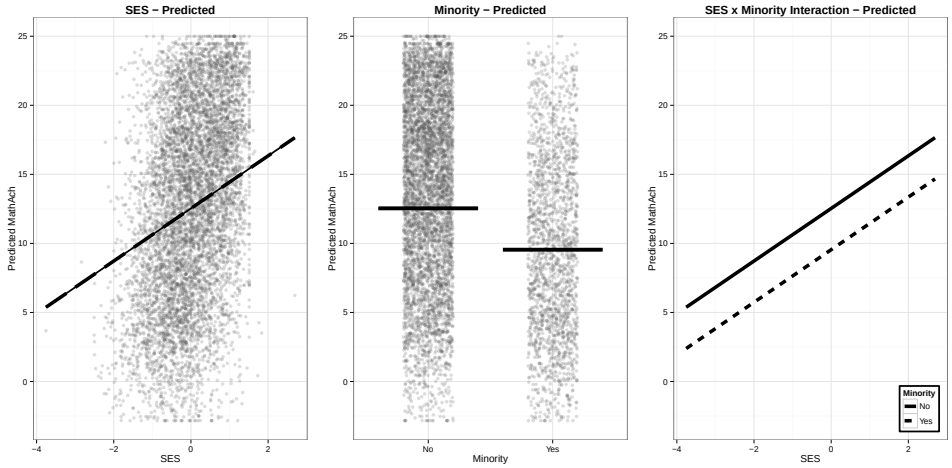


Figure 28: *Predicted value plots for SES and student minority status from a multilevel model. For SES, the actual prediction is shown with the solid line, while a smoothed loess approximation to this prediction is displayed with the dashed line. The interaction plot just depicts the smoothed approximation for simplicity, and shows no interaction because it was not specified. Each gray dot seen in both main effects plots represents an observation in the dataset.*

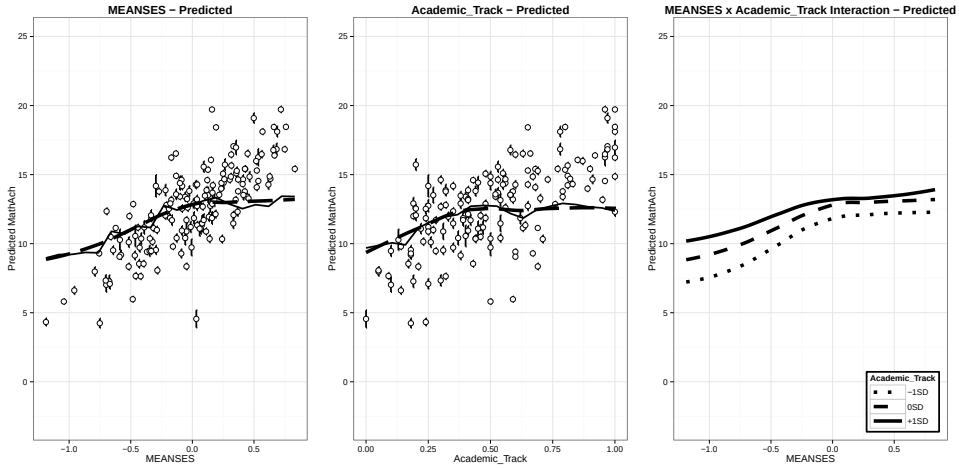


Figure 29: *Predicted value plots for the main effects and interaction between MEANSES and proportion of students on an academic track from a CART forest model. Note that because both measures are level-2 variables, results are aggregated and shown with error bars that represent the standard error of the mean for that particular school. The actual predictions for both variables are shown with the solid line, while a smoothed loess approximation to these predictions are displayed with the dashed line. The interaction plot just depicts the smoothed approximation for simplicity.*

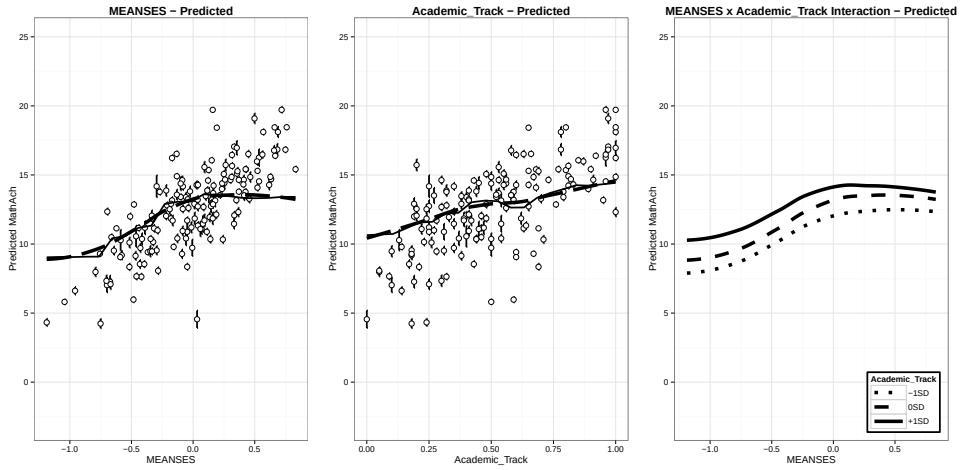


Figure 30: *Predicted value plots for the main effects and interaction between MEANSES and proportion of students on an academic track from a CFOREST model. Note that because both measures are level-2 variables, results are aggregated and shown with error bars that represent the standard error of the mean for that particular school. The actual predictions for both variables are shown with the solid line, while a smoothed loess approximation to these predictions are displayed with the dashed line. The interaction plot just depicts the smoothed approximation for simplicity.*

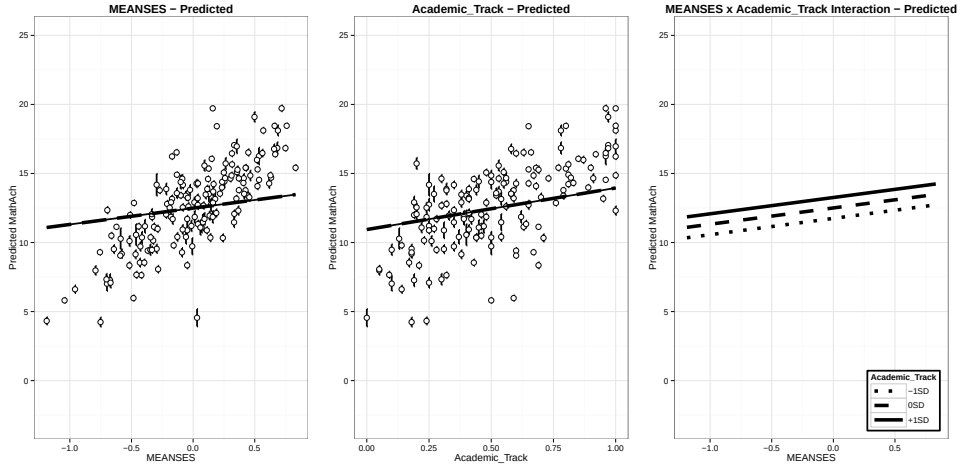


Figure 31: *Predicted value plots for the main effects and interaction between MEANSES and proportion of students on an academic track from a multilevel model. Note that because both measures are level-2 variables, results are aggregated and shown with error bars that represent the standard error of the mean for that particular school. The actual predictions for both variables are shown with the solid line, while a smoothed loess approximation to these predictions are displayed with the dashed line. The interaction plot shows no interaction because it was not specified.*

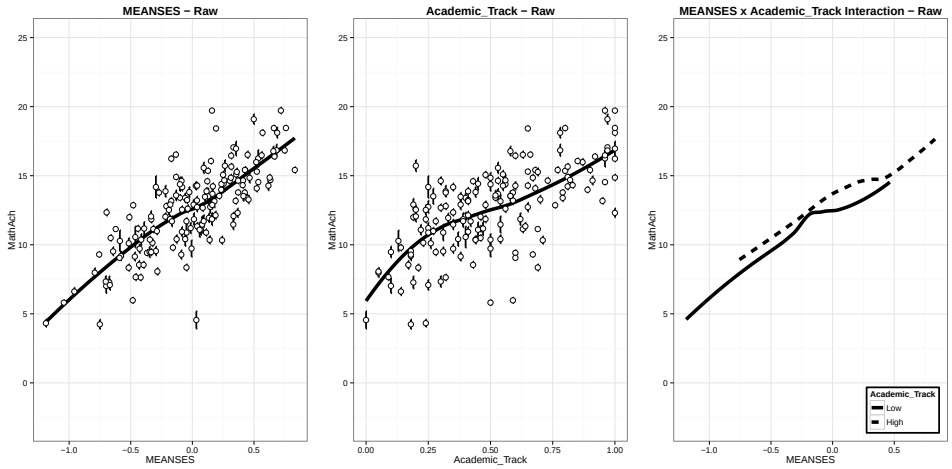


Figure 32: *Raw data for the main effects and interaction between MEANSES and proportion of students on an academic track using a loess smoother. Note that because both measures are level-2 variables, results are aggregated and shown with error bars that represent the standard error of the mean for that particular school. The actual predictions for both variables are shown with the solid line, while a smoothed loess approximation to these predictions are displayed with the dashed line. The interaction plot just depicts the smoothed approximation for simplicity.*

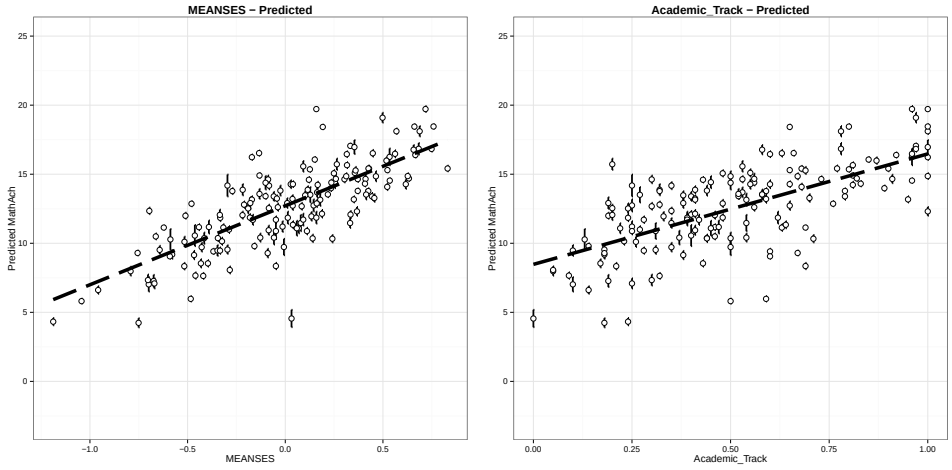


Figure 33: *The true predicted main effects for MEANSES and proportion of students on an academic track from a multilevel model. Note that these fits are much more accurate than the corresponding fits from the predicted value plotting procedure seen in Figure E, indicating that predicted value plots occasionally contain a loss of statistical information and results must be confirmed with true partial dependence plots and raw data plots.*

REFERENCES

Abry, T., Rimm-Kaufman, S. E., Larsen, R. A., & Brewer, A. J. (2013). The influence of fidelity of implementation on teacher–student interaction quality in the context of a randomized controlled trial of the responsive classroom approach. *Journal of School Psychology*, *51*(4), 437–453.

Adler, W., Brenning, A., Potapov, S., Schmid, M., & Lausen, B. (2011). Ensemble classification of paired data. *Computational Statistics & Data Analysis*, *55*(5), 1933–1941.

Adler, W., Potapov, S., & Lausen, B. (2011). Classification of repeated measurements data using tree-based ensemble methods. *Computational Statistics*, *26*(2), 355–369.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*(6045), 1034–1037.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using eigen and s4 [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=lme4 (R package version 1.0-5)

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*(3), 373–400.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York, NY: Springer.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

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

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Breiman, L. (2003). Manual–setting up, using and understanding random forests v4.0. Retrieved from http://oz.berkeley.edu/users/breiman

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

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, 927–961.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning* (pp. 161–168).

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Asso-*

ciation, *74*(368), 829–836.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J: L. Erlbaum Associates.

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2012). The value of replication for research on child development. *Developments: Newsletter of the Society for Research on Child Development*, *55*, 4–5.

Eliot, M., Ferguson, J., Reilly, M. P., & Foulkes, A. S. (2011). Ridge regression for longitudinal biomarker data. *The International Journal of Biostatistics*, *7*(1), 1–11.

Eo, S.-H., & Cho, H. (2013). Tree-structured mixed-effects regression modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, *23*(3), 740–760.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*(2), 275–297.

Fu, W., & Simonoff, J. S. (2015). Unbiased regression trees for longitudinal data. *Computational Statistics & Data Analysis*, *88*, 53–74.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, 1–5.

Goodnight, J. H. (1980). Tests of hypotheses in fixed effects linear models. *Communications in Statistics: Theory and Methods*, *9*(2), 167–180.

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development*, *82*(5), 1357–1371.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, *81*(4), 451–459.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328.

Hapfelmeier, A., Hothorn, T., & Ulm, K. (2012). Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. *Computational Statistics & Data Analysis*, *56*(6), 1552–1565.

Hapfelmeier, A., Hothorn, T., Ulm, K., & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, *24*(1), 21–34.

Hastie, T., & Tibshirani, R. (2013, December). *An interview with jerome friedman.* Retrieved from https://www.youtube.com/watch?v=79tR7BvYE6w

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan,

M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355–373.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013a). *An introduction to statistical learning*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013b). ISLR: Data for an introduction to statistical learning with applications in r [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=ISLR (R package version 1.0)

Karpievitch, Y. V., Hill, E. G., Leclerc, A. P., Dabney, A. R., & Almeida, J. S. (2009). An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. *PloS One*, *4*(9), e7087.

Knowles, J. E. (in press). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*.

Kopf, J., Augustin, T., & Strobl, C. (2013). The potential of model-based recursive partitioning in the social sciences: Revisiting ockham's razor. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (p. 75-95). New York, NY: Routledge.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, *12*(2), 361–386.

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*(4), 815–840.

Loh, W.-Y., Zheng, W., et al. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, *7*(1), 495–522.

Luke, D. A. (2004). *Multilevel modeling*. Sage.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86–92.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316.

Martin, D. P., & von Oertzen, T. (2015). Growth mixture models outperform simpler clustering algorithms when detecting longitudinal heterogeneity, even with small sample sizes. *Structural Equation Modeling*, *22*(2), 264–275.

McArdle, J. J. (2013). Exploratory data mining using decision trees in the behavioral sciences. In J. J. McArdle & G. Ritschard (Eds.),

Contemporary issues in exploratory data mining in the behavioral sciences (p. 3-47). New York, NY: Routledge.

Midgley, C., Maehr, M. L., Hruda, L. A., Anderman, E., Anderman, L., & Freeman, K. E. (2000). *Manual for the Patterns of Adaptive Learning Scale*. Ann Arbor: University of Michigan.

Mikami, A. Y., Boucher, M. A., & Humphreys, K. (2005). Prevention of peer rejection through a classroom-level intervention in middle school. *Journal of Primary Prevention*, *26*(1), 5–23.

Milborrow, S. (2014). plotmo: Plot a model's response while varying the values of the predictors [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=plotmo (R package version 1.3-3)

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*(1), 34.

Molnar, C. (2013). *Recursive partitioning by conditional inference* (Seminar Paper). LMU: Department of Statistics.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, *58*(302), 415–434.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18.

Open Science Collaboration. (2014). The reproducibility project: A model of large-scale collaboration for empirical research on reproducibility. In V. Stodden, F. Leisch, & R. Peng (Eds.), *Implementing reproducible computational research* (p. 299-323). New York, NY: Taylor & Francis.

Open Science Collaboration. (in press). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions.* New York, NY: Wiley.

Ottmar, E. R., Rimm-Kaufman, S. E., Berry, R. Q., & Larsen, R. A. A. (2013). Results from a randomized controlled trial: Does the responsive classroom approach impact the use of standards-based mathematics teaching practices? *Elementary School Journal*, *113*(3), 434–457.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Dubourg, V. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2015).

nlme: Linear and nonlinear mixed effects models [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=nlme (R package version 3.1-120)

Prindle, J. J., Brandmaier, A. M., McArdle, J. J., & Lindenberger, U. (2014). *Exploratory modeling with structural equation model forests.* (Poster presented at the 16th annual meeting of the Association for Psychological Science, San Francisco, CA.)

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Quinlan, J. R., & Cameron-Jones, R. M. (1995). Oversearching and layered search in empirical learning. In *Proceesings of the 14th international joint conference on artificial intelligence* (Vol. 2, pp. 1019–1024).

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77–93.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rice, J. A., & Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(1), 233–243.

Rieger, A., Hothorn, T., & Strobl, C. (2010). *Random forests with missing values in the covariates.* (Department of Statistics: Technical Reports, No. 79)

Rimm-Kaufman, S. E., Larsen, R., Baroody, A., Curby, T., Merritt, E., Abry, T., ... Ko, M. (2014). Efficacy of the responsive classroom approach: Results from a three year, longitudinal randomized control trial. *American Educational Research Journal*, *51*(3), 567-603.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Ruzek, E., Hafen, C., Allen, J., Gregory, A., Mikami, A. Y., & Pianta, R. (under review). How teacher emotional support motivates students: The mediating roles of peer relatedness and autonomy support.

Schelldorfer, J., Bühlmann, P., & van de Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using L1-penalization. *Scandinavian Journal of Statistics*, *38*(2), 197–214.

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations.* (Tech. Rep. No. NCEE 2008-4018). National Center for Education Evaluation and Regional Assistance.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, *87*(418), 407–418.

Segal, M. R. (1994). Representative curves for longitudinal data via regression trees. *Journal of Computational and Graphical Statistics*, *3*(2), 214–233.

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, *86*(2), 169–207.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Silberzahn, R. S., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (in prep). Many analysts, one dataset: Making transparent how variations in analytical choices affect results.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Strasser, H., & Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, *8*, 220-250.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348.

Therneau, T. M., & Atkinson, E. J. (2014). An introduction to recursive partitioning using the rpart routines. Retrieved from http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Tucker, L. R. (1966). Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology*. Rand McNally.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.: Addison-Wesley.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. Retrieved from http://www.jstatsoft.org/v45/i03/

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research.

Perspectives on Psychological Science, *7*(6), 632–638.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.