Efficiently exploring multilevel data with recursive partitioning

Daniel P. Martin & Timo von Oertzen

University of Virginia

Department of Psychology

Author Note

Special thanks should be given to Brian Nosek, Erik Ruzek, Micela Leis, Sara Rimm-Kaufman, and Vivian Wong for providing feedback on earlier versions of this work. Correspondence concerning this paper should be addressed to Daniel P. Martin, University of Virginia, dpmartin42@gmail.com. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B090002 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

There is an increasing number of datasets with many participants, many variables, or both, found in education and other areas of the social sciences that commonly have complex, multilevel data structures. Once initial confirmatory hypotheses are exhausted, it can be difficult to determine how best to explore these data sets 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. This paper argues that the utilization of a popular data mining framework known as recursive partitioning could offer 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.” The paper evaluates the feasibility of using these methods in multilevel contexts commonly found in social science research by using both Monte Carlo simulations and an applied example on student math achievement. Results indicate that while potential variable bias does occur in situations where variables are measured at multiple levels of analysis, these methods still prove to be a cost-effective way to revisit old datasets to discover something new in a data-driven way.

Keywords: random forests, data mining, exploratory data analysis

Efficiently exploring multilevel data with recursive partitioning

> 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 a 'mere descriptive statistics', no matter how much we had learned from it. (Tukey, 1977)

Consider the following scenario: An educational psychologist 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 in the social sciences 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 null hypothesis significance 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 child development (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 observations) 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 non-linear 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 paper 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. First, a brief introduction to recursive partitioning will be given, providing an 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. Following this, previous research will be discussed extending recursive partitioning frameworks to more advanced applications relevant to complex, multilevel data structures. The introduction will conclude with a set of unanswered questions when considering the application of these methods to multilevel data and includes a set of hypotheses based on previous research and statistical theory. Next, a simulation study will be outlined and the results will be presented and discussed. Using the information discovered in the simulation, a set of recommendations will be constructed for those interested in using these techniques. To further illustrate how these methods may be used in practice, an applied example using student math achievement from a subset of the High School and Beyond Survey will be shown. This paper will then conclude with a general discussion regarding the limitations and future directions of this research.

Recursive Partitioning

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 few decades, 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. As mentioned previously, this paper 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 briefly described below (for more information, refer to Strobl et al. (2009) for an accessible introduction to these methods). To assist in this overview, 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.

Classification and Regression Trees

Originally proposed by Breiman et al. (1984), CART is easily the most popular and widely used recursive partitioning algorithm. When 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 in the case of a continuous outcome or either entropy or the Gini index in the case of a categorical outcome.

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 residual sums of squares. 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.

Because decision trees that are too complex typically yield poor generalization performance, an additional step must be introduced into the process of building and evaluation a decision tree. This issue is handled 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. Typically, the amount of pruning corresponding to the best generalization performance is estimated using 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. This procedure is repeated K times, so every partition has an opportunity to act as a test set. The amount of pruning 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).

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.

Pros and Cons of CART. CART is a method that can efficiently search a parameter space, capturing potential non-linear 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, 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 10 classes will have 99 and 511 potential splits respectively, while a 5-point Likert scale item or a dichotomous variable will have substantially less (4 and 1, respectively).

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 only be explained conceptually for brevity. First, a global null hypothesis of independence between Y and all covariates X_j is tested. 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 strongest 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. Clearly, this algorithm is very similar to CART in its basic iterative partitioning structure. These two methods often yield similar predictive performance despite having different tree structures (Hothorn, Hornik, & Zeileis, 2006).

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 more computationally expensive compared to CART. Another major con with the conditional inference framework with respect to multilevel data is that the permutation test framework it is based on adheres to the traditional assumptions of independence.

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 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.

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).

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).

While random forests were originally based on the CART algorithm, the procedure is easily generalizable to any recursive partitioning algorithm. 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 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). Indeed, the predictive performance of both methods for the college graduation rate dataset were identical when rounded to the nearest whole number (MSE = 169).

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 observations left out in the bootstrap resampling process for each tree (i.e., the out-of-bag samples) are used to estimate predictive performance. Then, the same data are 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.

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 (Hastie et al., 2009). To calculate partial dependence for a given variable or variables, the entire training set must be utilized for every set of joint values in the variable(s) of interest.

As one can imagine, this can be quite computationally expensive for even datasets of moderate size. Luckily, a simplification can be made that results in vastly improved computation time with a minimal loss of statistical information. Rather than repeating the training set across all joint values of the variables of interest, 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. Repeating this observation not only results is a much faster computation time, but the predicted values are still easily interpreted 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 (Millborrow, 2014), they will be referred to as *predicted value plots* here for clarity.

Pros and Cons of Random Forests. 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, partial dependence plots, and predicted value plots all 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.

Extending Recursive Partitioning to Multilevel Data

There is a small, but growing body of research that examines the application of decision trees to complex 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).

Other methods focused on applying recursive partitioning to longitudinal data employed 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, et al. (2013) also utilized GUIDE for this purpose.

More recently, researchers have also begun to focus on estimating and using 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 have been extended to 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.

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 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. 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. In the social sciences, however, it is common to include variables measured at different levels of the analysis.

Additionally, not one of the studies that was previously mentioned have actually examined the original non-parametric methods outlined in the previous section in multilevel contexts commonly found in the social sciences. Because of this, the performance of these methods with regard to prediction and variable importance accuracy are not well understood. Below, potential issues are discussed that manifest themselves conceptually when considering the application of the CART and CTREE algorithm to multilevel data structures.

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 in a theoretical sense, 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 should be selected more often purely due to chance.

Multilevel Issues with Conditional Inference. As mentioned previously, the conditional inference algorithm is based on a permutation test framework and 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 data sets 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 removed in the aggregation procedure 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.

With the quantitative issues outlined above, the behavior of recursive partitioning methods need to be better understood in multilevel contexts before they can be used as exploratory data analytic tools in education research. The goal of this simulation study is to evaluate the performance of recursive partitioning methods, specifically forest methods, in situations commonly found in social scientific research (specifically education) 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 the simulation study 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.

Method

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

Methodological Approach

Random forests using classification and regression trees (CART forest). CART forests were run using randomForest, version 4.6-10 (Liaw & Weiner, 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. Variable importance was extracted from the fitted model object, which is based in a standard permutation framework.

Random forests using conditional inference trees (CFOREST). Conditional inference forests were run using party, version 1.0-20 (Strobl, Boulesteix, Zeileis, & Hothorn, 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 (the default). Each tree was built using sub-sampling rather than bootstrap re-sampling as suggested by Strobl et al. (2007) for unbiased variable importance. Variable importance was extracted from the fitted model object, which is based in a permutation framework as described by Hapfelmeier, Hothorn, Strobl & Ulm (2014).

Multilevel regression. Multilevel regression models were run using lme4, version 1.1-7 (Bates, Maechler, Bolker, & Walker, 2014). 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. Variable importance was only calculated for the main-effects model, and thus was simply the p-value for each respective variable (calculated using a Satterthwaite approximation for the denominator degrees of freedom).

Simulation Conditions

In total, five parameters were of primary interest to be manipulated in a 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 manipulated parameter 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 cross-sectional research in education 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 cross-sectional educational research, while an ICC of 0.30 is meant to represent a large ICC level found in cross-sectional educational research.

Finally, the level at which the covariates are measured will also be manipulated. Given that preliminary simulations showed potential issues with level-2 variables, the variables will be simulated to be either level-1 only, or both level-1 and level-2. A condition with only level-2 variables is not needed, as such an analysis typically aggregates the outcome to the second level, creating a new data set that is no longer multilevel in nature. Thus, with a fully crossed simulation design, there are 24 (4 x 3 x 2) simulation conditions. Each condition was run 200 times for a total of 4800 datasets. See Table 1 for a summary of the main parameters that were manipulated in this study.

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 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 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 was simulated. The model equations can be seen below, with the models 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, 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:

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X1_{ij}) + \beta_{2j}(X1_{ij}^2) + \beta_{3j}(X2_{ij}) + \beta_{4j}(X1_{ij}X2_{ij}) + r_{ij}$$

Level 2:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \mu_{0j}\\ \beta_{1j} &= \gamma_{10} + \mu_{1j}\\ \beta_{2j} &= \gamma_{20}\\ \beta_{3j} &= \gamma_{30}\\ \beta_{4j} &= \gamma_{40}\end{aligned}$$

Both levels:

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X1_{ij}) + \beta_{2j}(X1_{ij}^2) + r_{ij}$$

Level 2:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01}(X2_j) + \mu_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(X2_j) + \mu_{1j}\end{aligned}$$

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 (i.e., 0.1) and medium (i.e., 0.3) 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.

Evaluation Criteria

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

Proportion variation explained. This metric is based on mean squared error, which is a common metric that is used 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 - MSE/var(y)). Note that if the predictions are very bad, it is possible to have negative values. Because the true multi-level 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. 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 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 hold out test set.

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 importance was calculated via the built-in permutation-based procedure for both forest models and by the corresponding p-values for each variable for the mixed-effects model (calculated using a Satterthwaite approximation for the denominator degrees of freedom).

Simulation results

First, simulation results will be presented for the proportion of variance explained metric. Next, simulation results for variable importance will be presented, aggregated by each condition manipulation.

Variation results

As expected, the naive main-effects only multilevel model performed the worst with respect to predictive ability, followed 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 predictors were included at both levels of the analysis. Sample size also played a large role in prediction, where estimates of the proportion of variation explained metric became more stable in conditions with larger sample sizes. In fact, negative values for proportion of variation accounted for 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. Overall, both CART forests and CFOREST perform similarly. See Figure 2 for the results of the proportion variation accounted for metric by method, ICC, covariate level, and sample size.

Variable importance aggregated across conditions

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. Conditional forests are close to this desired pattern, which is expected given they are based on an unbiased algorithm. These 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 forest showed its well-documented bias toward variables that were continuous (variables 1, 3, 4, and 5) or non-binary (9 and 10). Variable 2, the categorical variable with a simulated relationship with the outcome, was even overlooked by the CART forest in favor of continuous variables with null relationships with the outcome. Overall, sample size did not seem to play a large role for the variable importance metric like it did with the proportion variation accounted for metric. See Figure 3 for the results of the variable importance metric by variable, statistical method, and sample size.

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, CFOREST yields unbiased variable importance measures when the covariate level is level-1 only, regardless of ICC. Bias is introduced in CFOREST when covariates are included at both levels of the analysis, and this bias is apparent even when the ICC is negligible. Somewhat surprisingly, CART forests show an artificial preference for variables included at the second level of analysis, and like CFOREST, this bias increased as the ICC increased. See Figure 4 for the results of the variable importance metric by variable, statistical method, ICC, and covariate level.

Simulation-based Guidelines

Based on the results of this simulation, the following recommendations are proposed for applied researchers interested in using these techniques for either exploration or prediction. First, use the ICC from an intercept-only model in conjunction with the levels at which your variables were measured to determine any potential biases that may arise when running a forest model. Then, run both a random forest and a conditional inference forest to estimate the proportion of variation accounted for (in the case of a continuous variable) or misclassification 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).

Next, compare this value to one derived from a naive, main-effects only multilevel model. If the forest methods perform worse than the naive model, then there is little evidence to support any nonlinearities or interactions in the dataset. However, if the forest methods perform similarly or better than the naive model, then there is evidence for potential nonlinearities or interactions, and these should be investigated through the visual inspection of variable importance plots and partial dependence plots (keeping in mind of potential biases, of course). Despite the bias found in this simulation, forest methods can still provide a useful avenue for exploration and prediction in multilevel designs as long as researchers are wary of what kind of bias is introduced. The following example will show how these methods might be used in practice.

Application

High School & Beyond is a nationally representative survey of U.S. public and Catholic high schools conducted by the National Center for Education Statistics. 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). The R script to replicate these results can be found in the supplemental materials.

Measures

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.

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.

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.).

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.

Methodological approach

Step 1: ICC. Running an intercept-only model 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.

Step 2: Estimate proportion of variation explained. Given the larger sample size, the data were split into a training set 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 2 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 unlikely to generalize to a future sample.

Step 3: Examine variable importance and predicted value plots. Both variable importance plots and predicted value plots were examined. Variable importance can be seen in Figure 5. 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 variables, both forest methods actually chose two 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.

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 6. 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. 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 ($R^2_{\text{level-1}} = 0.003$)

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. Both the predictive performance comparison and further investigation with a follow-up model found the magnitude of this interaction effect to be extremely small. In total, there was not much evidence for specifying more complex functional forms beyond a main-effects only model for this dataset.

General Discussion

The main goal of this paper 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 provide a brief outline of how these techniques might be used in practice. With regard to the simulation study, there were three main findings of note.

First, with regard to conditional inference forests, 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 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. Research is now beginning to focus on extending the RE-EM algorithm mentioned previously to conditional inference trees (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. As this simulation showed, however, simply having an ICC of zero is not enough to remove all the bias in variable importance when covariates are at both levels. Incorporating the nesting into the permutation framework through some sort of multilevel permutation or post-hoc correction in the conditional inference algorithm could be a potential solution to be investigated.

Second, and most surprising, was that the bias found in variable importance for CART-based algorithms. Despite have a simulated relationship with the outcome via a small main effect and a medium interaction effect, variable 2 was not detected as having a strong relationship with the outcome given the fact that it was a binary variable. 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 analysis. This is especially true when the sample sizes were large. This finding indicates that these methods could serve as effective prediction tools in multilevel designs when theory might not be able to dictate how variables are related with respect to potential nonlinearities or higher order interactions and ICC values are reasonable (or all variables occur at the first level).

Limitations and Future Directions

Like with any simulation research, the findings 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 longitudinal research likely would require different parameter values. A simulation based on a longitudinal study, for example, would require 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 absence of missingness in the simulation study. Missing data are an all-too-common occurrence in many areas of psychology and education, especially with respect to complex, multilevel designs. While not investigated due to brevity, recursive partitioning methods have a few elegant ways to handle missing data. CART forests employ an imputation method that iteratively builds forest models 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).

Conditional inference forests, on the other hand, rely on surrogate splits. 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 et al., 2006). 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 decision tree methods under missingness conditions that are completely at random (Hapfelmeier, Hothorn, & Ulm, 2012). One important thing to note is that both of these methods only work when predictors are missing. Observations with the outcome missing must either be removed or imputed with a separate process. Future research could investigate how these methods perform in the context of multilevel data under a variety of missingness mechanisms, with both covariates and the outcome potentially missing.

Finally, note that recursive partitioning was specifically chosen as the focus of this paper 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 predictive 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 applied dataset example, 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 multilevel framework (e.g., Eliot, Ferguson, Reilly, & Foulkes, 2011; Schelldorfer, Bühlmann, & van de Geer, 2011), such a model is quite complex and unlikely to be learned and adopted for the sole purpose of exploratory data analysis. However, if the focus is purely 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 whether the models accounts 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 techniques, and a naive main effects only model at various deviations from linearity to see which methods perform the best.

Conclusion

This paper examined new ways to conduct exploratory research on multilevel datasets in the social sciences using recursive partitioning. Both simulation and an applied example have shown that these methods can 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.

References

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 1.1-7)

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

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

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

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*, 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, 30,* 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).

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*, 1–11.

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

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

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation, 84*, 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*, 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*, 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*, 355–373.

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

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*, 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*, e7087.

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

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

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

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

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

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.

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)

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, *58*, 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, 615–631.

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.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 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 1.1-20)

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

Quinlan, J. R., & Cameron-Jones, R. M. (1995). Oversearching and layered search in empirical learning. In *Proceedings 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/.

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

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*, 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*, 407–418.

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

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

Silberzahn, R. S., et al. (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*, 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,* 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*, 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*, 323–348.

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

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*, 632–638.

Table 1

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
Covariate level	2	level-1 only, both levels

Table 2

Proportion of variation in math achievement explained by each method in the High School and Beyond Survey.

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

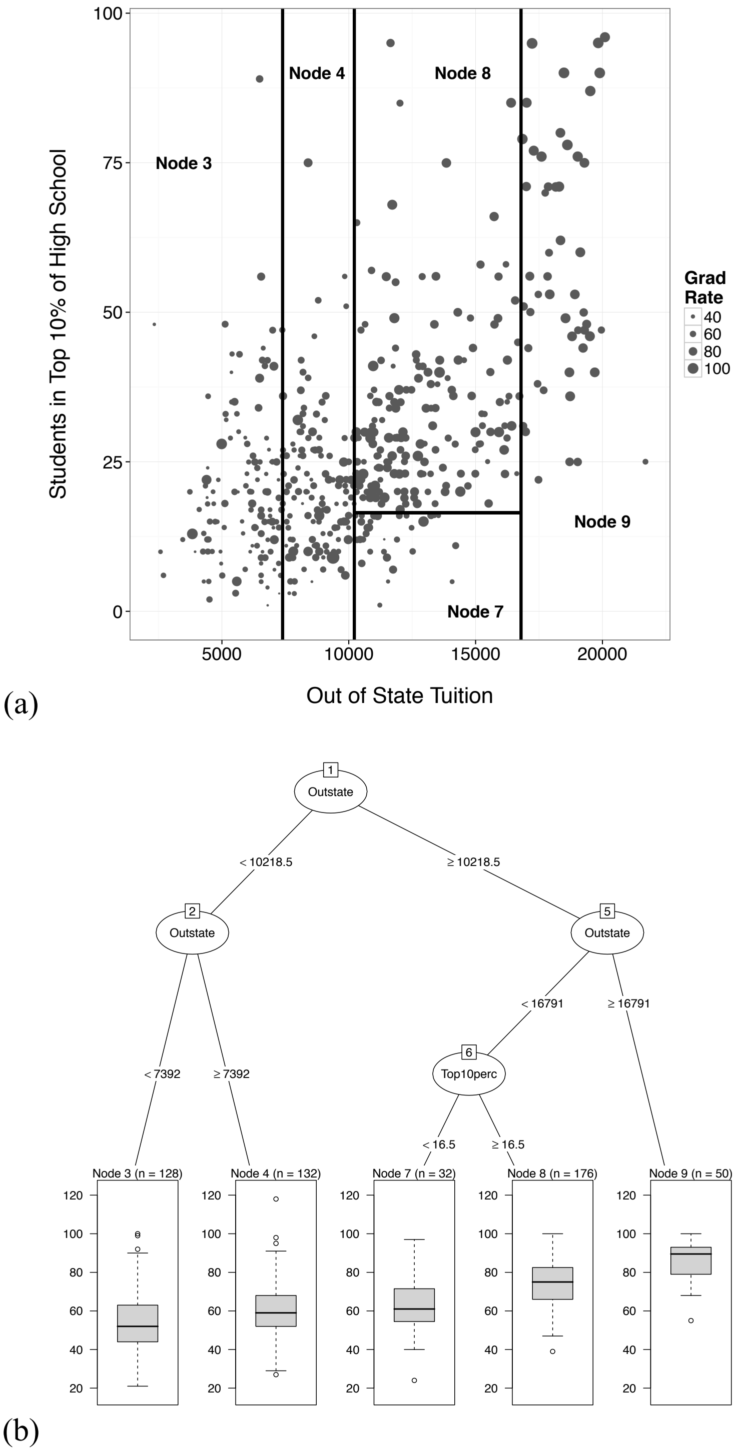

Figure 1. (a) How college graduation rates vary as a function of out-of-state tuition and percentage of students in the top 10% of their high school class. These partitions are created by a set of binary decision rules, shown in (b).

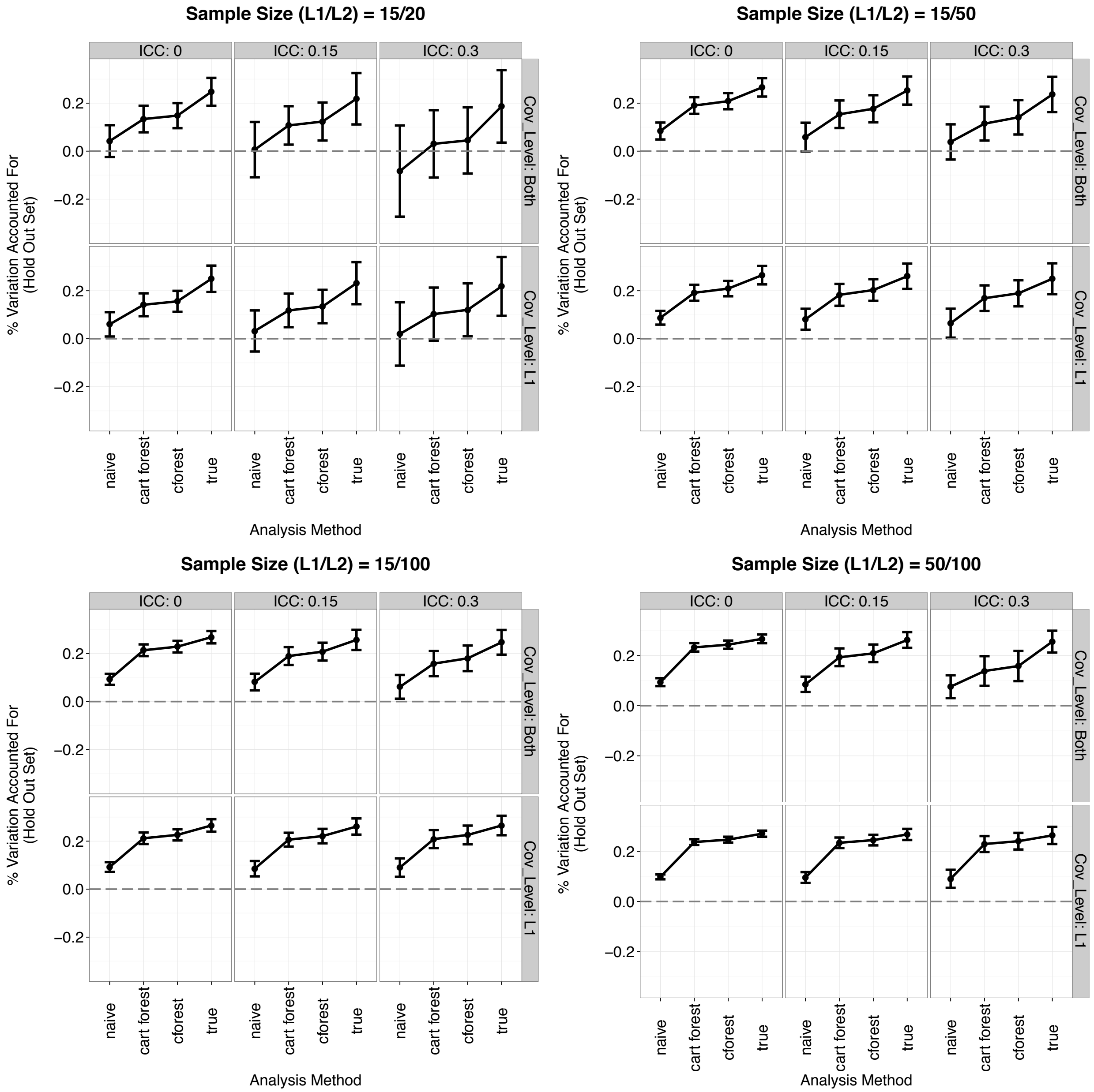

Figure 2. Proportion variation estimated metric by statistical method (x-axis), ICC value (columns), and covariate level (rows). Each matrix depicts the same plot, but with different sample sizes. Error bars depict the standard deviation of 200 simulations within each condition.

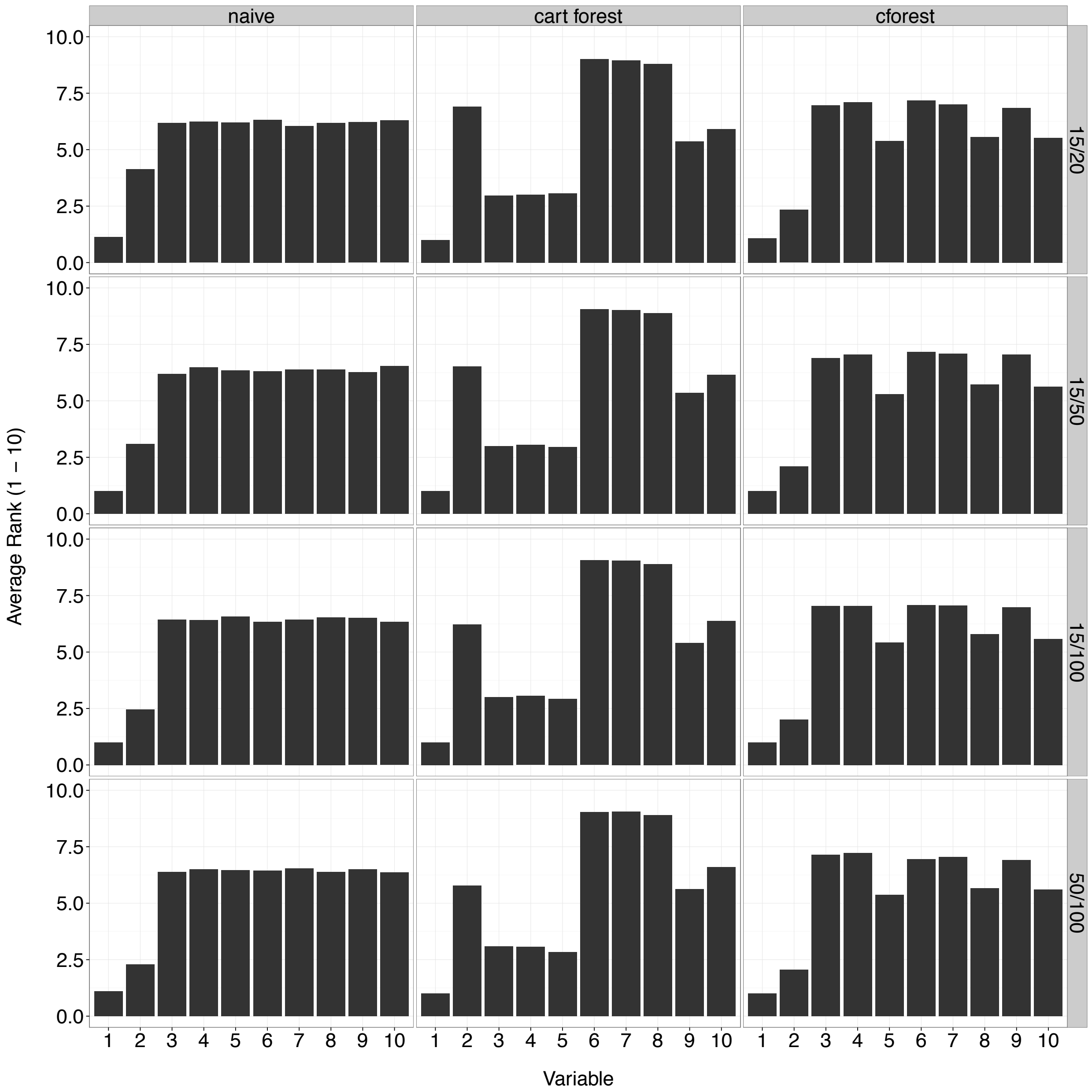

Figure 3. Variable importance metric by variable (x-axis), statistical method (column) and sample size (row). Ranks were created and averaged across the 200 simulations within each condition.

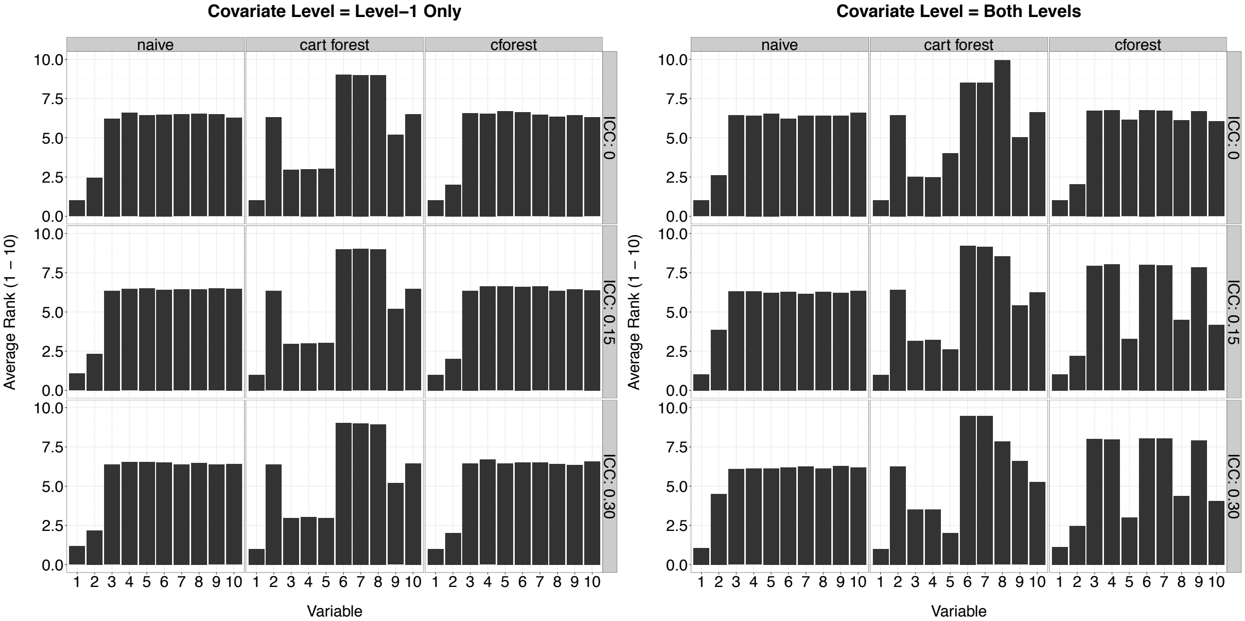

Figure 4. Variable importance metric by variable (x-axis), statistical method (column), ICC (row), and covariate level (facet). Ranks were created and averaged across the 200 simulations within each condition.

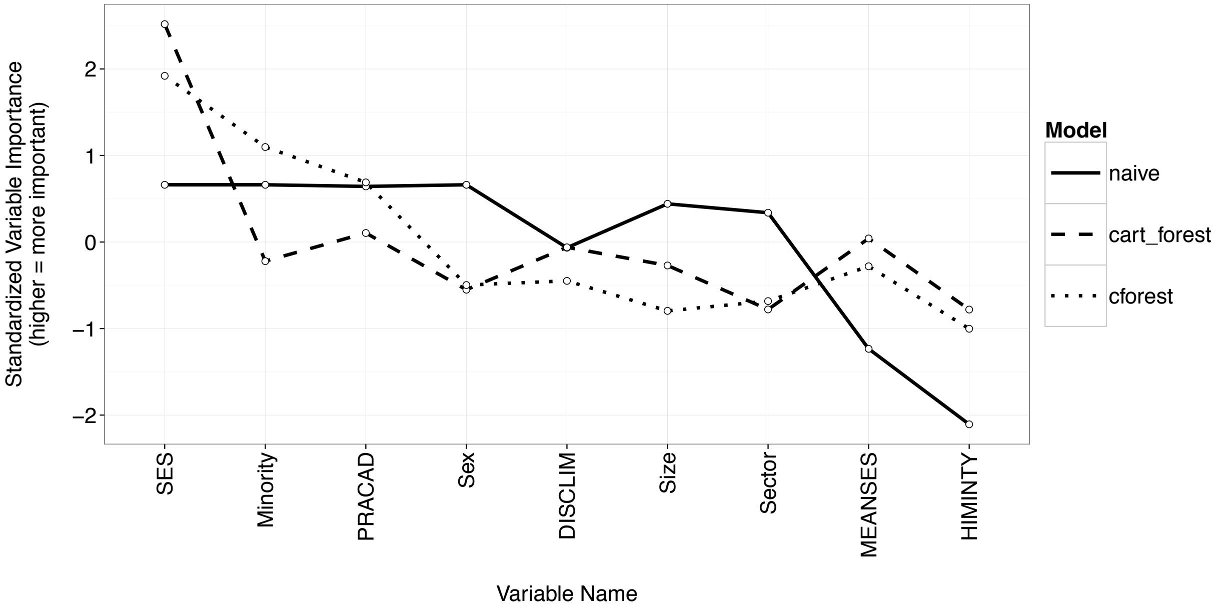

Figure 5. Variable importance for a naive main-effects only model, a cart forest, and cforest. Note that importance was standardized to allow for easier comparisons. The x-axis is ordered such that the most important variable (averaged across all models) is the furthest left.

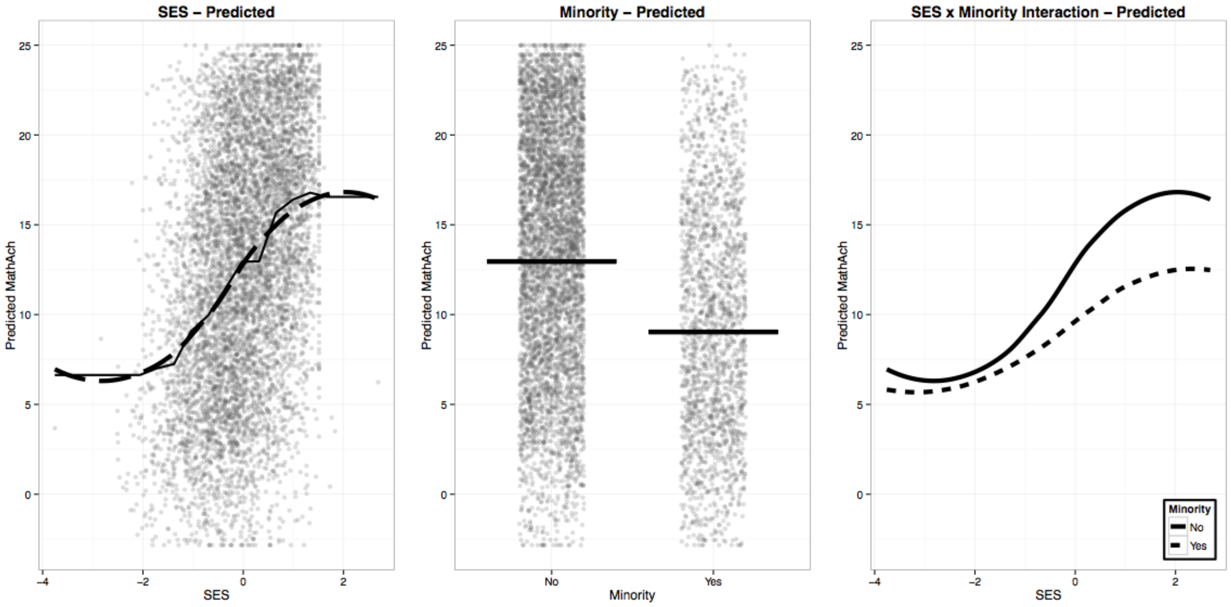

Figure 6. Predicted value plots for SES and student minority status from a CFOREST model. For SES, the actual prediction is shown with the solid line, while a smoothed loess approximation to this prediction is displayed with the dashed line. The interaction plot just depicts the smoothed approximation for simplicity. Each gray dot seen in both main effects plots represents an observation in the dataset.