Abstract Title Page

Title: Efficiently Exploring Multilevel Data with Recursive Partitioning

Authors and Affiliations:

Daniel P. Martin & Timo von Oertzen
University of Virginia, Department of Psychology

Sara E. Rimm-Kaufman
University of Virginia, Curry School of Education

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

There is an increasing number of datasets with many participants, variables, or both, in education and other fields that often deal with large, multilevel data structures. Keeping with the topic of the spring 2015 SREE conference, the term "multilevel" in this context can refer to either cross-sectional data structures, such as children nested within classrooms, or longitudinal data structures, such as repeated-measures nested within participants. Once initial confirmatory hypotheses are exhausted, it can be difficult to determine how best to explore the dataset to discover hidden relationships that could help to inform future research. Naturally, this practice is often done "by hand." That is, the researcher in question will run multiple tests with different combinations of predictors to help identify important variables related to the outcome.

From an implementation perspective, this approach can be quite difficult and time-consuming as every analysis will require different analytic considerations, such as potential variable transformations, methods of handling missing data, different software packages to use, etc. Additionally, even though each analysis might be grounded in scientific theory, this approach is still rife with statistical issues, such as blurring the lines between exploratory and confirmatory research (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) and capitalizing on chance discoveries resulting in an increased number false positive results (Nelson, Simmons, & Simonsohn, 2011), among others.

What is needed, then, is a method that can: 1) efficiently search a large parameter space of potential predictor variables while controlling for the potential for detecting false positives; 2) output a variable importance metric to identify potential variables related to the outcome that might have been overlooked in previous research; 3) adequately describe potential non-linear relationships or higher-order interactions that might not have been considered without explicit specification by the user; and 4) handle missing data in a straightforward way. These four desirable characteristics can all be accomplished with a non-parametric data mining approach called Random Forests.

Purpose / Objective / Research Question / Focus of Study:

While recent research has begun to investigate Random Forests in a multilevel context (e.g., Hajjem, Bellavance, & Laroque, 2011; Sela & Simonoff, 2012) this research tends to focus solely on predictive accuracy and fails to provide solutions to common problems that would arise if this method were to be applied to an educational domain. As such, the purpose of this study is to examine the feasibility of applying Random Forests to efficiently explore large, multilevel datasets commonly found in educational research.

Significance / Novelty of study:

The majority of research examining recursive partitioning methods in the presence of multilevel data has been performed in the last few years in Data Mining community focusing on predictive accuracy. For example, both Sela & Simonoff (2012) and Hajjem, Bellavance, & Laroque (2011) independently proposed the same method to incorporate a given random effects structure in a recursive 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 to encompass an entire tree. Because of the ability to add random effects after filtering through the tree structure, both methods show increased predictive accuracy compared to either a traditional multilevel model or a decision tree without random effects.

However, while these recently developed methods are certainly improvements, they do have issues. For example, the algorithm these methods are based on is biased toward categorical variables with many levels and continuous variables with wide ranges (Hothorn, Hornik, & Zeileis, 2006). Because of this, variable importance measures are biased towards these variables (Strobl, Boulesteix, Zeileis, & Hothorn, 2007), leading to erroneous discoveries. Additionally, these methods have no method of calculating variable importance when missing data are present (Hapfelmeier, Hothorn, Ulm, & Strobl, 2014), which commonly occurs in educational research. This study looks to examine the performance of Random Forests and potential corrections to these problems that were recently proposed (Hapfelmeier et al, 2014), focusing specifically on multilevel contexts.

Statistical, Measurement, or Econometric Model:

A Random Forest is a popular, non-parametric data mining method that creates ensembles of simple decision trees (see Breiman, 2001 and Strobl, Malley, & Tutz, 2009, for a review). A decision tree is a method that recursively partitions the predictor space to create homogenous sub-groups based on a given outcome, which can be either categorical (resulting in a classification tree) or continuous (resulting in a regression tree). Splitting rules depend entirely on the specific algorithm chosen. The following steps outlines the general algorithm:

1) Search all variables for potential splits
2) Identify the best split by some criterion
3) Split the sample on this threshold, resulting in two child nodes
4) Repeat step 1-3 on the resulting sub-groups (nodes) until reaching a stopping criterion

To help illustrate a simple decision tree in practice, I will be using an educational dataset (N = 160) with math achievement as the outcome that is publicly available using the *nlme* package in R. To keep this example simple and avoid the multilevel structure of the data, the data was aggregated to the school level. This resulted in eight variables: percentage of students who are in a minority group (isMinority), percentage of students who are male (isMale), school size (Size), whether the school was a Catholic school (isCatholic), percentage of students on an "academic track" (PRACAD), a measure of discrimination climate (DISCLIM), and the average socio-economic status of the school (MEANSES). Note that while it is important to run the same tree on a separate test set or employ some kind of cross-validation technique to avoid overfitting, a discussion of these methodological issues is beyond the scope of this introduction.

See Figure 1 for the decision tree. The first split occurs on mean socio-economic status, with schools with values lower than -0.284 filter to the left child node, while schools with values higher than -0.284 filter to the right child node. This binary decision process then continues through each inner node, until the final split is reached, resulting in a terminal node. As evident in the graph, school socio-economic status, as well as percentage of students on an "academic track," seem to be related the most with overall school math achievement.

A Random Forest, then, is just a logical extension of a simple decision tree. The general algorithm follows the following steps:

1) Draw a bootstrap sample (i.e., sample with replacement) from the original dataset
2) Draw a sample of m (typically the square root of the total number of variables) from the original dataset
3) Create a decision tree
4) Repeat steps 1-3 to create a specified number of trees (typically around 500)

New observations are then filtered through each individual tree, and the resulting predictions are aggregated together to create on overall prediction. Because this method creates many trees by taking a bootstrap sample of observations and a sample of variables, Random Forests are less likely to make mistakes commonly found in decision trees, such as being susceptible to minor shifts in the data, or having only a few important variables mask the effects of others. However, while Random Forests are better predictors than simple decision trees, they inevitably become less interpretable due to the large number of single trees that are aggregated together. Thus, this poster will highlight two main techniques to help with the interpretation of Random Forests: variable importance and partial dependence plots.

Variable importance refers to a value assigned to each variable that determines how important that variable is to predicting the outcome. Creating such a measure involves a simple, three-step process. First, the predictive accuracy of a dataset is calculated using the aggregation mentioned above. Then, the same dataset 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 predictive accuracy and the permuted predictive accuracy. If the difference is large, then this indicates that the particular variable plays an important role in predicting the outcome. Otherwise, if the predictive accuracy does not change very much, this variable is concluded to have no relationship with the outcome. See Figure 2 for the variable importance plot for the math achievement example. This plot seems to confirm what was found in the simple decision tree: socio-economic status and percentage of students on an "academic track" seem to be the most important for predicting math achievement.

Partial dependence plots are graphical visualizations of the marginal effect of a given variable (or multiple variables) on an outcome. While partial dependence plots are useful for help examine main effects and potential interactions, they are restricted to only two or three variables due to the limits of human perception. See Figure 3 for an example of such a plot, which only includes the top five predictors discovered from the Random Forest. As evidenced in the plot, all five predictors seem to follow a linear pattern.

Usefulness / Applicability of Method:

This method has the potential to increase the ability for researchers to perform efficient exploratory data analysis without the common pitfalls and methodological challenges that were mentioned earlier. It is important to note that because this method is non-parametric, there is no need to explicitly model random effects to account for nesting (as there are no standard errors). However, the algorithm does need to be slightly altered in its splitting and re-sampling procedure to ensure error estimates are accurate. The poster to be presented at SREE will both provide the necessary correction for this technique to be applied to multilevel contexts and will apply the technique to a simulated dataset. This application will be also annotated with sample code and include a referral to a walkthrough existing on the first author's website with the goal of giving applied researchers the necessary information needed to use this tool effectively in their own multilevel research.

Conclusions:

While more in-depth simulation work is currently underway to investigate the use of Random Forest at varying levels of non-independence, it seems that this method is both a feasible and relatively easy-to-understand statistical tool for applied researchers to effectively explore their data to help uncover potential hidden relationships and identify variables that might have been overlooked in the confirmatory hypothesis testing phase. Additionally, consistent with the theme of the spring 2015 SREE conference, these methods are not simply limited to cross-sectional multilevel data structures, but can also be applied to longitudinal (i.e., repeated-measures nested within individuals) to identify potential non-linear trajectories in individual growth trajectories as well as detect potential interactions between growth patterns and additional covariates.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

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

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

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.

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.

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.

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.

Appendix B. Tables and Figures

Not included in page count.

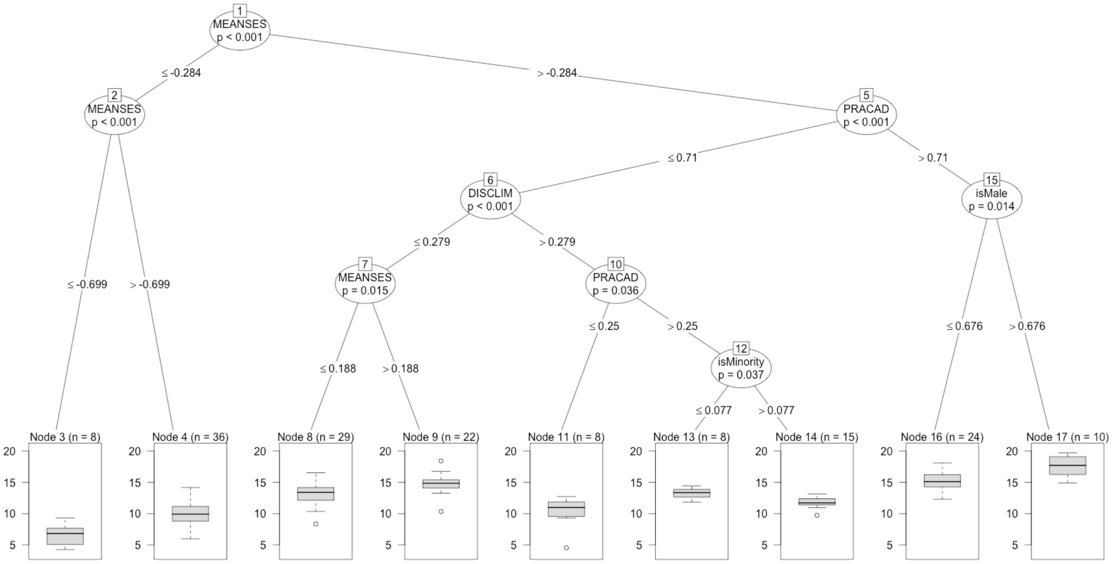

Figure 1. A simple decision tree of the math achievement dataset aggregated to the school level. Box plots reflect the distribution of each terminal node.

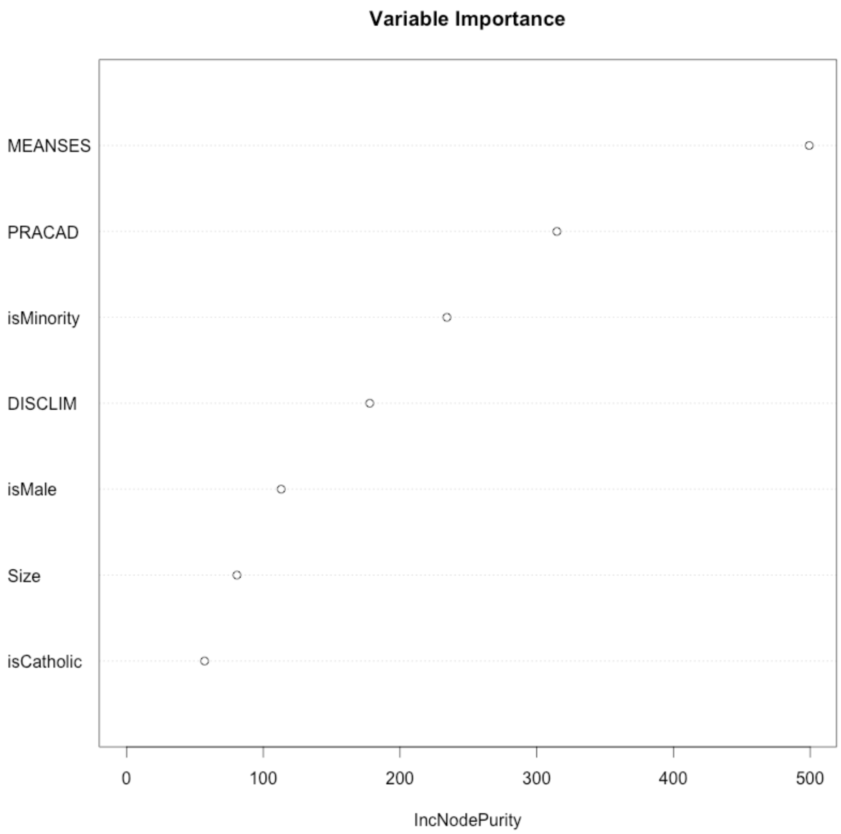

Figure 2. A Variable Importance plot for a Random Forest using the math achievement dataset. Larger values correspond to higher variable importance.

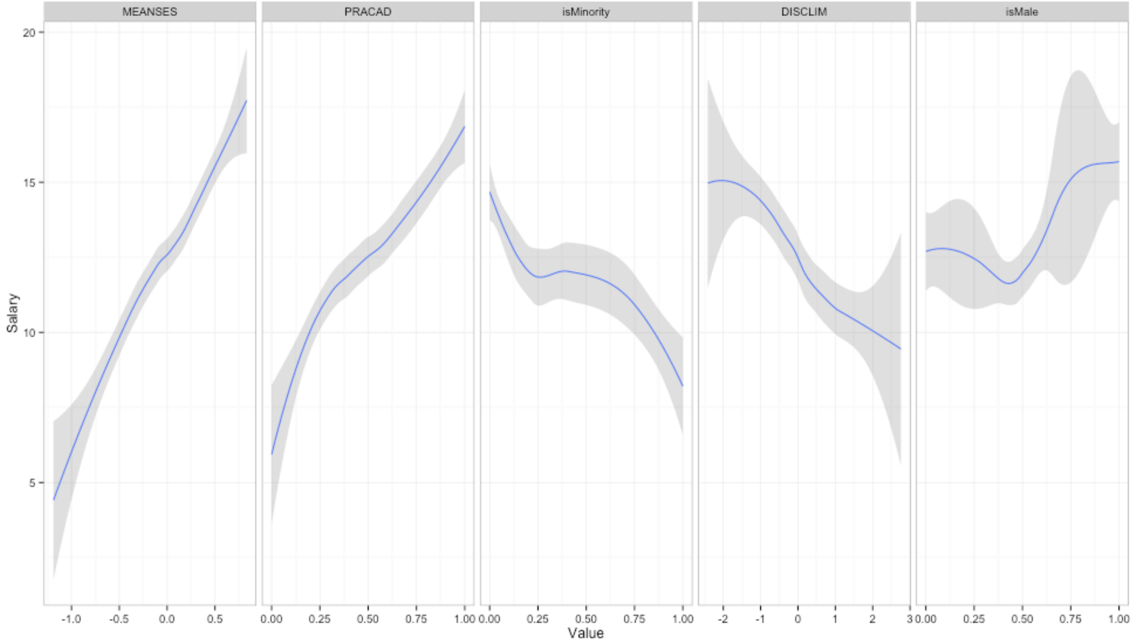

Figure 3. A partial dependence plot for the top five predictors discovered in the Random Forest for the math achievement data.