Recursive Partitioning in the Presence of Multilevel Data

W. Holmes Finch
Ball State University

Researchers in the social and behavioral sciences are increasingly working with data that are sampled at multiple levels, where individuals at the first level are nested within clusters at the second level, and in some cases these level-2 clusters are nested within clusters at a third level. Examples of this type of data structure can be found in large scale data programs such as the National Assessment of Educational Progress (NAEP), the Program for International Student Assessment (PISA), and Trends in International Mathematics and Science Study (TIMMS), among others. Such data require the use of special data analysis strategies that appropriately account for the presence of the multilevel data structure, in order to avoid parameter estimation bias. Simultaneous with this increase in research being done using multilevel models, has been the use of data mining techniques to fully explore relationships among variables in large and complex data files. Of these approaches, one of the most popular involves recursive partitioning methods such as classification and regression trees, and random forests. Until very recently, multilevel versions of these partitioning models were not available, leading to difficulties for researchers working with multilevel data structures who want to use recursive partitioning. This study showcases two such methods for recursive partitioning in the multilevel data context. Detailed analyses are conducted, results of these analyses are discussed, and implications of the models are described. The R code used to carry out these analyses is provided in the appendix to the manuscript.

Multilevel data, in which individuals are nested within clusters, is very common in educational and psychological research (Gumus, 2014; Van Laere, Aessaert, & van Braak, 2014; Winnaar, Frempong, & Blignaut, 2015). Research involving such data might involve a research design in which schools are the primary sampling unit, and all students within the selected schools participate in the study. Likewise, similar designs are commonly employed in the health sciences, where patients are nested within hospitals, and in psychology, where clients are nested within therapists, as examples. In addition to such planned experiments, multilevel data also appears in research scenarios involving large datasets, such as the Programme for International Student Assessment (PISA), the Early Childhood Longitudinal Study (ECLS), and the National Assessment of Educational Progress (NAEP), to name just a few. In each of these cases, individuals are nested within schools, which are in turn nested within a larger organizing entity such as a state or a nation.

Coincident with the increased use of multilevel data in educational and psychological research, has been the greater popularity of very flexible modeling techniques based upon recursive partitioning of data in order to identify predictor variables that are related to an outcome variable of interest. Such models have become increasingly popular for both prediction of dependent variable values, and exploration of variable importance because they do not rely on the common assumptions underlying standard linear regression models, including multivariate normality, and homogeneity of variance (Holden, Finch & Kelley, 2011; Lei & Koehly. 2003; Pai, Lawrence, Klimberg & Lawrence, 2012; Rausch & Kelley, 2009; Finch & Schneider, 2007). Given their flexibility with regard to the distributions underlying both the response and predictor variables, as well as the ability to automatically detect nonlinear relationships among the variables, such recursive partitioning models are more frequently used in the social sciences (e.g., Gruenewald, Mroczek, & Ryff, 2008; Markham, Young, & Doran, 2013; Ozgen, Hellemann, & de Jonge, 2013). However, relatively little work has been done in regards to the use of such models in the presence of multilevel data. Therefore, the purpose of the current study was to demonstrate two proven methods for fitting recursive partitioning models with multilevel data. Each of these approaches combines the flexibility of recursive partitioning techniques with the appropriate modeling of multilevel data structure. Following is a brief review of multilevel modeling, recursive partitioning, and ensemble recursive partitioning. Next, prior research describing problems associated with multilevel data and recursive partitioning models is presented, followed by a discussion of multilevel methods for recursive partitioning. Finally, the data that serves as the motivation for this study are presented, followed by a detailed description of the study results and a discussion of their implications, particularly with respect to the use of the methods described here.

Multilevel Models

Multilevel models (MLMs), sometimes also referred to as mixed effects models, are used in the analysis of data in which individuals (level-1) are nested within clusters (level-2), and the clusters could themselves be nested within higher order clusters (level-3). MLMs frequently occur in educational research, where individual students are sampled by (and therefore nested within) schools. They also

appear in psychological research, where clients receiving therapy are nested within therapist, and in medical research where individuals patients may be nested in doctors and hospitals. In all of these cases, the modeling of an outcome variable must account for the nested structure of the data in order to ensure that standard errors and model parameters are accurately estimated (Snijders & Bosker, 2012). One of the most common such MLMs is the random intercept model linking an independent variable, *x*, with a dependent variable *y*. This model takes the form:

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij} \quad (1)$$

where; y_{ij} =Dependent variable value for individual *i* in cluster *j*
β_{0j} =Intercept for cluster *j*
β_1 =Slope relating independent variable *x* to dependent variable *y*
x_{ij} =Value of *x* for individual *i* in cluster *j*
ε_{ij} =Random error for individual *i* in cluster *j*

In model (1), β_{0j} can be expressed as:

$$\beta_{0j} = \gamma_{00} + U_{0j} \quad (2)$$

where; γ_{00} =Mean intercept across clusters
U_{0j} =Unique effect of cluster *j* on the intercept

The parameter γ_{00} is what is known as a fixed effect, meaning that it takes the same value for all clusters. On the other hand, U_{0j} is a random effect that varies across clusters. In the context of students nested within schools, this would mean that model intercepts would differ across schools, with part of the intercept including a common component across schools (γ_{00}), as well as a component unique to the individual school (U_{0j}). In model (1), β_1 is a fixed effect and is therefore constant across clusters. Again, in the school research context, this would mean that the relationship between the independent and dependent variables is the same for all schools. It is also possible to fit a random coefficients model in which β_1 has both fixed and random components, just as is true here for β_{0j}. The model parameters in (1) and (2) are typically estimated by maximum likelihood (ML) or restricted ML (REML) estimation, which differ in terms of how the standard errors of parameter estimates are calculated. Specifically, the degrees of freedom used in ML does not account for the fact that the parameters themselves are being estimated, leading to a negative bias in the standard error estimates (Kreft & de Leeuw, 1998). In contrast, REML standard error estimates do use degrees of freedom that account for the estimation of the model parameters, thereby producing unbiased estimates. REML is the more commonly used of the two approaches because its standard errors are generally more accurate (Kreft & de Leeuw).

Classification and Regression Trees

A potential drawback of linear models, including the MLM described above, is that it restricts the relationships between predictors and response to be linear and additive in nature (no interactions among predictors) unless the researcher explicitly includes interaction terms. If the researcher is not aware of important interactions in the population, or unsure of the form that they might take, then such terms will not be included in the analysis, resulting in a misspecified model. One set of methods for regression that can seamlessly include all manner of nonlinear relationships without the researcher having to prespecify anything about the nature of the model except the predictor variables are those based on recursive partitioning (RP). Classification and regression trees (CART) is a very popular RP method that was first outlined by Breiman, Friedman, Olshen, & Stone (1984). CART develops a prediction model for an outcome (*Y*) given a set of independent variables (*x*) by iteratively dividing individual members of the sample into ever more homogeneous groups, or nodes, based on values of the predictor variables. It is a nonparametric method in that there are no assumptions regarding the functional form of the model linking the dependent and independent variables (Williams, Lee, Fisher & Dickerman, 1999). Thus, it is not limited to fitting linear, or even additive relationships. In addition, it can very easily incorporate variables of different types and measured on different scales (Strobl, Malley, & Tutz, 2009). Thus, for example, nominal categorical variables with more than two categories can be included in the analysis without dummy coding and the attendant selection of a referent category that is necessary when using such variables with standard regression models, or MLM. In addition, CART is able to fit nonlinear

relationships between the predictors and the outcome, with the added advantage of being able to deal with interactions automatically. Following is a basic description of the general CART approach, along with discussion of limitations to the method. It should be noted that there are a variety of approaches for dealing with the particulars of CART, and that the discussion below is designed to present the general methodology.

In order to understand the manner in which CART works, let us consider an example in which each member of the sample has a score on a reading test, and each individual is measured on a set of continuous and categorical predictor variables. CART begins by placing all subjects into a single root node at the top of the tree. It then searches the entire set of predictors to find the value for one of them by which it can divide the observations into two new nodes whose values on *Y* are as homogeneous as possible. In this example the outcome variable of interest is the continuous variable examinee reading score. Thus, the optimal split is the one resulting in two nodes with the lowest variance possible. To continue the example, the predictors include 5 subtests of a diagnostic assessment designed to provide schools and teachers with information regarding various reading subskills, including Vocabulary (sub1), Informational text structures (sub2), Informational text comprehension (sub3), Literary text structures (sub4), and Literary text comprehension (sub5). The CART algorithm assessed every possible split for each of these subtests and found that splitting the sample at a score of 202 on subtest 5 (literary comprehension) yielded the most homogeneous daughter nodes. Thus, all individuals in the sample with a literary comprehension score less than or equal to 202 were moved to the left side of the tree, while those with scores greater than 202 were moved to the right side. The exact determination of which variable to split where can be made in a number of ways, including minimizing a loss function comparing the observed and CART predicted values on the outcome variable, or calculating a Chi-square test statistic to determine whether the split results in significantly different patterns for the outcome variable categories in the resulting daughter nodes.

For each daughter node, the predictors are once again searched for the optimal split by which the subjects can be further divided into ever less heterogeneous nodes. This division of the data continues until a predetermined stopping point is reached such that further splits do not appreciably reduce the heterogeneity of the resulting nodes. For a continuous outcome variable, within node heterogeneity can be expressed by the deviance (*D*) statistic:

$$D_i = -2\sum n_{ik} \ln S_i^2 \tag{3}$$

where; n_{ik} = Number of subject from group *k* in node *i*
S^2_i = Variance of the dependent variable in node *i*.

Larger values of D_i indicate greater heterogeneity within the node; i.e. a relatively less optimal solution. The sum of these individual node deviances,

$$D = \sum D_i \tag{4}$$

is a measure of the overall performance of the CART solution, with smaller values indicating greater within node homogeneity and thus a better performing tree. A number of potential stopping rules have been proposed for CART, such as stopping when a minimum terminal node size (e.g. 5) has been reached, when no additional splits result in a statistically significant Chi-square statistic, or when a particular threshold of D_i for each of the potential terminal nodes has been achieved.

The predicted value of the dependent variable yielded by the CART model for an individual is the mean value of *Y* for the terminal node where that individual is placed. Variable importance is measured by the decrease in D that comes from using the variable in a particular split. Thus, in the context of CART, variable importance is directly tied to specific splits within the tree, rather than being a global measure across all branches of the tree. The interested reader can trace other parts of the tree to learn more about the relationships among the subtest scores and reading proficiency.

CART has a number of advantages that make it an attractive alternative to traditional linear models. However, it also has some distinct disadvantages that can prove problematic to the researcher. Perhaps foremost among these problems is the tendency to overfit the training sample, due to the relative sensitivity of the method to sample specific characteristics (Hastie, Tibshirani, & Friedman, 2009). For

example, decisions regarding on which variables to split and where the splits should occur have been shown to be closely tied to the distributional characteristics of the training sample and therefore may not generalize well to the broader population (Bühlmann & Yu, 2002). The result can be a CART model that fits the training data quite nicely but does not provide particularly accurate predictions for other samples. One early approach for dealing with this problem was the process of pruning, whereby branches in the tree are removed if they do not reduce the overall heterogeneity of the model by a predetermined amount (Ripley, 1996). While frequently yielding superior trees to the original in terms of classification accuracy for a second (cross-validation) sample, pruning is somewhat cumbersome to carry out, and may not totally eliminate the problem of overfitting (Hothorn, Hornik, & Zeileis, 2006). More modern methods of growing classification trees, such as the approach used in this study relying on the Chi-square test, do not require pruning. Given this combination of possessing several advantages vis-à-vis other modeling methods, while at the same time having the potential for overfitting the training data, an alternative approach to classification that relies on trees while overcoming some of their distinct problems would be most desirable, and indeed has been developed in the form of the ensemble recursive partitioning methods, Bagging and Random Forests.

Ensemble Recursive Partitioning Models

While CART can be overly sensitive to characteristics in the training sample and thereby will not always produce a generalizable solution, it is also important to note that predictions from a single CART analysis are unbiased so that averaged over a number of individual trees the resulting predictions for an individual should be quite accurate (Bauer & Kohavi, 1999; Dietterich, 2000). Based upon this fact, researchers have extended CART with alternative methods for developing predictive models based upon the recursive tree model outlined above. These two methods, Bagging (Brieman, 1996) and Random Forests (RF; Brieman, 2001) each rely on bootstrap resampling to overcome the aforementioned problems with CART. Both Bagging and RF select a large (e.g., 1000) number of *B* bootstrap samples and apply CART to each of these, yielding *B* individual trees. These bootstrap samples can either be drawn with replacement and be the same size as the original or without replacement and represent subsets of the original sample. Each bootstrapped tree yields a predicted value for members of the sample, and the results of the *B* trees are then averaged to ascertain both variable importance information, and to predict an individual's group membership. The difference in the two ensemble methods is that Bagging makes use of the entire set of predictors for each tree, and RF applies bootstrapping to the predictors as well as to the sample. Thus, for each RF tree *B* bootstrap samples of subjects and predictor variables are used. Individuals not included in bootstrap sample *B* are referred to as the out of bag (OOB) sample for that particular tree. Because the trees used by RF are even more diverse than those used in Bagging, it can be shown that its averaged results are also less sensitive to sample specific variation and thus more generalizable (Brieman, 2001). In addition, by relying on bootstrapped samples of predictor variables, RF is able to yield more information than Bagging or CART regarding the true importance of all predictor variables. This is because some of the RF trees will not include very dominant variables that mask the importance of other predictors, giving these less important predictors an opportunity to demonstrate their contribution to the prediction. Given these advantages, RF will be the method of ensemble prediction of primary interest in this study.

With regard prediction for a continuous outcome variable using ensemble methods, the set of trees that have been fit in the model building process are applied to each new individual for whom a predicted outcome is desired. Thus, each tree is applied to each individual subject as was described for CART, and the final predicted value on *Y* for an individual in each such application is recorded. After all of the trees have been applied, each individual receives as their predicted value on *Y* the mean across the *B* RF trees.

In addition to prediction, determining variable importance is also frequently an important issue to researchers. In the context of MLM this is typically done using hypothesis tests of model parameters associated with individual variables. When variables are significantly related to the outcome variable, they are deemed to be important. Variable importance in RF can be calculated using a permutation methodology (Nicodemus, Malley, Strobl, & Ziegler, 2010). This approach works by permuting one or more variables so as to eliminate naturally occurring relationships in the data (Edgington & Onghena, 2007). Typically, a large number of such permutations are made in order to create a distribution of permuted outcomes. Next, the statistic of interest from the original dataset is compared to the

permutation distribution. For RF, the permutation importance of an individual predictor variable is calculated by comparing the number of correct predictions made by the actual data (i.e. the predictor ordered as it appears in the original dataset) with the number of correct predictions made when the variable has been permuted (i.e. randomly shuffled), averaged across all trees in the ensemble. Thus, for example, in order to obtain the permutation variable importance for subtest 1, we would randomly reorder the subtest 1 scores across the subjects, create a tree, and obtain predicted outcomes for each subject. This permuting of subtest 1 would be carried out a large number of times (e.g. 1000). Then, we would compare the prediction accuracy across trees for the original variable with that of the mean for the permuted trees. If the difference is large, and in favor of the tree based on the original data, we would conclude that the variable is important in accurately predicting the outcome variable. On the other hand, if the difference in prediction accuracy between the actual and permuted values is very small, then we would conclude that the variable does not contribute much more to predicting *Y* than if it were random and thus totally unrelated to the outcome. These variable importance values can then be compared graphically, as will appear later, in order to determine which variables are most important in accurately predicting group membership. More formally, importance for variable x_m for a single tree (*t*) is calculated as:

$$VI_t(x_m) = \frac{\sum I(y_i=\hat{y}_{iO})}{|\bar{B}|} - \frac{\sum I(y_i=\hat{y}_{iP})}{|\bar{B}|} \quad (5)$$

where; $\hat{y}_{iO} =$ Prediction for observed data
$\hat{y}_{iP} =$ Prediction for permuted data
B=out-of-bag (OOB) sample

If variable x_m is not included in the tree, then *VI*=0. In order to obtain the overall variable importance measure for the RF, we then calculate

$$VI(X_m) = \frac{\sum_{t=1}^{T} VI_t(x_m)}{T} \quad (6)$$

where *T* is the total number of trees in the ensemble.

Recursive Partitioning with Multilevel Data

There is a growing body of research examining the impact of multilevel data on the performance of recursive partitioning models such as CART and RF. Results of simulation studies have demonstrated that certain aspects of recursive partitioning algorithm performance are deleteriously impacted by multilevel data, whereas other aspects appear not to be affected at all. For example, in a simulation study, Karpievitch, Hill, Leclerc, Dabney, and Almeida (2009) showed that RF produced equally accurate predictions of an outcome variable, whether data were multilevel or single level. Similar results regarding prediction accuracy with multilevel data have also been shown for CART (Fu & Simonoff, 2015; Strobl, Malley, & Tutz 2007). In short, what research exists examining the impact of multilevel data on RF and CART prediction accuracy suggests that these methods will perform as well as with multilevel data as when the data are single level only.

Despite these positive results regarding prediction accuracy, however, there does appear to be some impact of multilevel data on the calculation of variable importance measures for recursive partitioning methods. For example, conditional inference methods for recursive partitioning have at their core the assumption of independence of errors for the individual observations. When this assumption is violated due to clustering of individuals, the test used to determine whether a split should be made is biased, and this bias increases as the Intraclass Correlation (ICC), which measures the relationship of scores for individuals within the same cluster, increases (Luke, 2004). Bias in this statistic leads to the potential for identification of more splits in the sample tree than is actually true in the population, and in turn identification of more variables as being important than should be the case. Karpievitch, et al. (2009) found that for multilevel data, the individual trees that make up the forest in RF are highly correlated with one another, and that this correlation increases concomitantly with increases in the ICC. In turn the inflated correlation among the trees results in an underestimate of the OOB error. As noted above in equation (6), the OOB sample is used to calculate variable importance statistics. Thus, underestimation of OOB error can result in biased estimates of variable importance, leading to inaccurate conclusions regarding which of the predictor variables are most and least important in characterizing the outcome variable. Finally, research has shown that in general, CART is more likely to use predictors with more

values when selecting variables for splitting the data (Loh & Shih, 1997). Thus, when level-1 and level-2 continuous predictors are both included in the analysis, level-1 variables are more likely to be used in splitting because they can take N possible values as opposed to level-2 variables which can take K possible values, where N and K represent the number of level-1 (e.g. people) and level-2 (e.g. schools) units, respectively. Given that they are more likely to be used in splitting, level-1 variables will therefore be reported by CART as more important than level-2 variables, as a rule, even when such is not the case in the population (Martin & van Oertzen, 2015).

Given the potential difficulties with using standard recursive partitioning models with multilevel data, particularly in terms of identifying important variables, two alternative approaches have been suggested. These methods, Mulilevel Exploratory Data Analysis (MLEDA), and Mixed Effects Mulilevel Recursive Partitioning Trees (RE-EM) take very different approaches to accounting for multilevel data structure when using recursive partitioning models. However, both methods are designed to correct some of the aforementioned problems created by multilevel data. Each of these approaches is described below in brief, with more references provided for readers more interested in an in depth technical understanding of these methods.

Multilevel Exploratory Data Analysis (MLEDA)

MLEDA (Martin, 2015) represents an extension of the way in which variable importance measures are calculated for CART and RF, rather than as a new algorithm for growing trees. As noted above, multilevel data does not appear to unduly influence predictions obtained from RP methods. However, calculation of variable importance for these approaches, particularly RF, is impacted by multilevel data structure because the trees are correlated with one another, leading to correlated OOB samples and errors. In order to counter this problem and thereby obtain more accurate estimates of variable importance, Martin proposed using a simulated cross-validation sample for calculating the statistic (6), rather than the OOB sample. This simulated sample is generated using marginal traits from the data (i.e. means and variances appearing in the data), as well as the covariance structure that is found in the data. Using this information, more accurate variable importance measures that are not influenced by the cluster correlated nature of the data can be constructed, and should more accurately identify important variables, and order them appropriately. Simulation research (Martin) has shown that indeed the methodology for calculating variable importance based upon the simulated cross-validation sample yielded more accurate ordering of variable importance than did the traditional approaches for this purpose.

Random Effects Expectation Minimization Recursive Partitioning (RE-EM Tree)

Sela and Simonoff (2012) have proposed a very different solution to the problem of fitting recursive partitioning algorithms (in particular CART) to multilevel data. Martin's (2015) method for dealing with this issue focused on correcting the calculation of the variable importance statistics, and did not involve changing the basic algorithm used to grow the tree. In contrast, Sela and Simonoff developed a new method algorithm for fitting trees with multilevel data, by estimating the random effect(s) in the model and removing them prior to actually growing the tree. Put another way, the RE-EM tree approach models the effects of the data being clustered, and then removes or conditions on them when fitting the tree. When an initial tree has been fit, the random effects are estimated again. This cycle is continued until convergence is reached. Following is a more formalized presentation of the RE-EM tree algorithm emphasizing its various steps.

1. Set the estimated random effects to 0 (i.e. assume that there is no effect associated with cluster membership).
2. Fit CART to the data in the standard way predicting dependent variable *y* using the predictors, and record terminal node membership of each individual in the sample.
3. Estimate the mixed effects model in equation (1) using the dependent variable *y* and the predictors.
4. Retain the estimated random effects.
5. Calculate $y_i - Z_i b_z$, where Z_i is the value of the random effect for person *i*, and b_z is the random effect coefficient for effect *z*.
6. Iterate through steps 2-5 until convergence is reached for b_z.
7. Replace the predicted values of y produced by CART with the mean of y_i obtained by the multilevel model for individuals in each terminal node.

The RE-EM tree method is designed to appropriately account for multilevel data in the construction of the tree, and estimation of the predicted values obtained from the tree. While Sela and Simonoff (2012) focused on a random intercepts model, it is possible to incorporate any random effects model into the RE-EM tree framework, including models with random slopes for one or more of the predictors. However, it should also be noted that random effects are not used in the actual construction of the tree, meaning that variables with random slopes will not be used to in the recursive partitioning process, but only in the estimation of the random effects themselves, and the associated adjustment of *y* in step 5 above.

Goals of the Current Study

The primary goal of this study was to demonstrate the use of the two recursive partitioning methods for multilevel data that have been described above. Each method is demonstrated using a multilevel dataset, and the results are interpreted in some detail. It is hoped that researchers faced with multilevel data and interested in using RP might find this manuscript helpful in developing an analysis plan of their own, based upon either of these methods. The appendix includes the R code used in this study.

Methods

Data for this study came from the Progarmme for International Student Assessment (PISA). The sample included a total of 10,000 randomly selected 15-year old students from the 65 nations participating in the Programme for International Student Assessment (OECD, 2009). The outcome variable of interest in this study was the math achievement test score obtained as a part of PISA. The independent variables for these analyses were individual student SES, the mean of student SES at the school level, and the mean of student SES at the country level. The PISA SES index is derived from a factor analysis of variables that include parent education and occupation and possessions in the home (mean = 0, standard deviation = 1). The means at the school and country levels represented the "typical" SES for individual schools and countries. In addition to SES, student sex was also included as an independent variable, as were scores on metacognition and learning strategy inventories. PISA (OECD) included two metacognitive indexes and three learning strategy use indexes. Metacognition was measured as metacognitive knowledge about text comprehension. Students were presented with scenarios and then evaluated the quality and usefulness of strategies for reaching an intended goal. The rank order of the strategies was compared to an optimal ranking developed by experts. Two metacognitive indexes were created: The index of understanding and remembering (UNDREM) and the index of summarizing (METASUM). Additionally there were three learning strategy indices: The frequency of use of control strategies (CSTRAT), memorization strategies (MEMOR), and elaboration strategies (ELAB). All of these variables were expressed as *z* scores. Two multilevel recursive partitioning methods were used to fit the data, including MLEDA and RE-EM tree. Results from each method are described below, with particular emphasis on the identification of important variables for differentiating students based upon math test performance. The R code used to carry out these analyses appears in the appendix.

Results

Results of the data analyses are organized by method, followed by a general summary of what was found.

MLEDA

When using MLEDA, the first step in data analysis is to estimate the ICC, which is done by fitting the random intercept only model (including no predictor variables) in the standard multilevel modeling context. The ICC is the amount of variance in math test scores associated with country (3295) divided by the total variance in math test scores (10093), or 0.326. Thus, just under 33% of the variation in math test scores is associated with the nation in which the examinee lives. In other words, a third of the variability in math scores across individuals is associated with their home country. The presence of such a large ICC suggests that clustering is an important facet of this study, and that multilevel appropriate methods for fitting recursive partitioning models should be strongly considered. If the ICC were very small, then we would not anticipate a strong effect due to the clustering, so that a naïve method for fitting the trees would probably work fine.

The RF MLEDA model was fit to the data such that math achievement was the outcome, with the predictors being all of those listed in the methods section. The proportion of variance in math explained by the RF model, based upon a cross-validation sample was 0.229. In other words, a model was fit to a training set of individuals made up of randomly selected 50% of the total sample, and then the model was applied to the other 50%, who made up the cross-validation sample. For this second sample, the RF model accounted for 22.9% of the variance in math test scores.

Of particular interest in the current study was the determination as to which variables were most important in predicting math achievement. As noted above, variable importance measures can be particularly susceptible to bias in the context of multilevel data. Figure 1 includes variable importance values for each of the predictors, for the RF model and a naïve model in which no interactions of the variables were included. The naive model results are included as a standard component of MLEDA graphics output and serve as a baseline against which the RP results can be compared. The degree of divergence that is seen in results for the two models indicates the extent to which including interactions through the use of RP is helpful for understanding the dependent variable. If the variable importance results for the two methods are quite similar, then there is no advantage to using the more complex RP model. Results in Figure 1 do show a great deal of divergence between the two approaches, however. In particular, the RF MLEDA model reveals that the most important predictors of math achievement are all associated with relative wealth, with school SES, followed by student SES, and country SES having the highest importance values. In contrast, the learning strategies are somewhat less important in predicting math performance, with the most salient being the metacognitive strategies of summarizing and understanding/remembering. The least important factor in terms of math achievement was student sex.

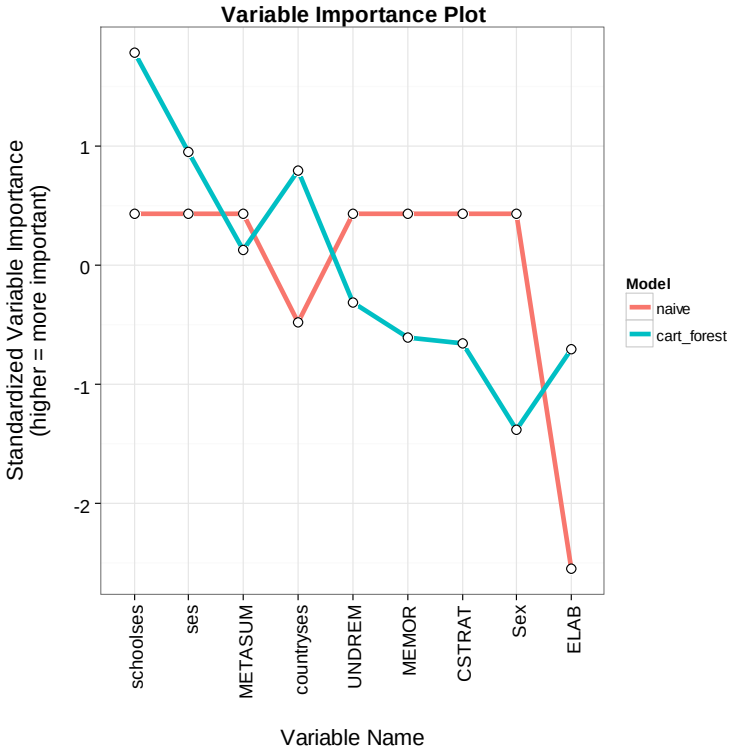

Figure 1. Variable importance plots for the naïve and RF models with math achievement score as the outcome.

In addition to the relative importance of each variable in terms of predicting the outcome, it is also of some interest to ascertain the nature of the relationships between the outcome and predictor variables, which can be done using predicted value plots for the RF model. Figure 2 includes such plots for the predictors. Each panel in the figure represents the relationship between the model predicted math test score (*y*-axis) and a predictor variable (*x*-axis). This plot is particularly helpful for understanding how the outcome is related to each of the predictors. For example, we can see that math scores increase concomitantly with increases in the three SES values. In addition, this increase in math performance appears to accelerate somewhat for higher school SES, SES, and country SES values, before leveling off at the very highest levels of wealth. Finally, math scores were also positively related to the use of METASUM, and UNDREM.

The predicted value plots can be used to investigate possible interactions in the data as well as the main effects. For example, Figure 3 includes a set of three graphs for the interaction of student SES and METASUM with regard to model predicted math test performance. The first graph in the triptych shows the relationship between SES and predicted math score only, with the second displaying the relationship between METASUM and predicted math score. The third graph in this set includes predicted math score on the *y*-axis and student SES on the *x*-axis, with separate lines for 1 standard deviation below the mean, the mean, and 1 standard deviation above the mean on METASUM. In this way, we can see that the

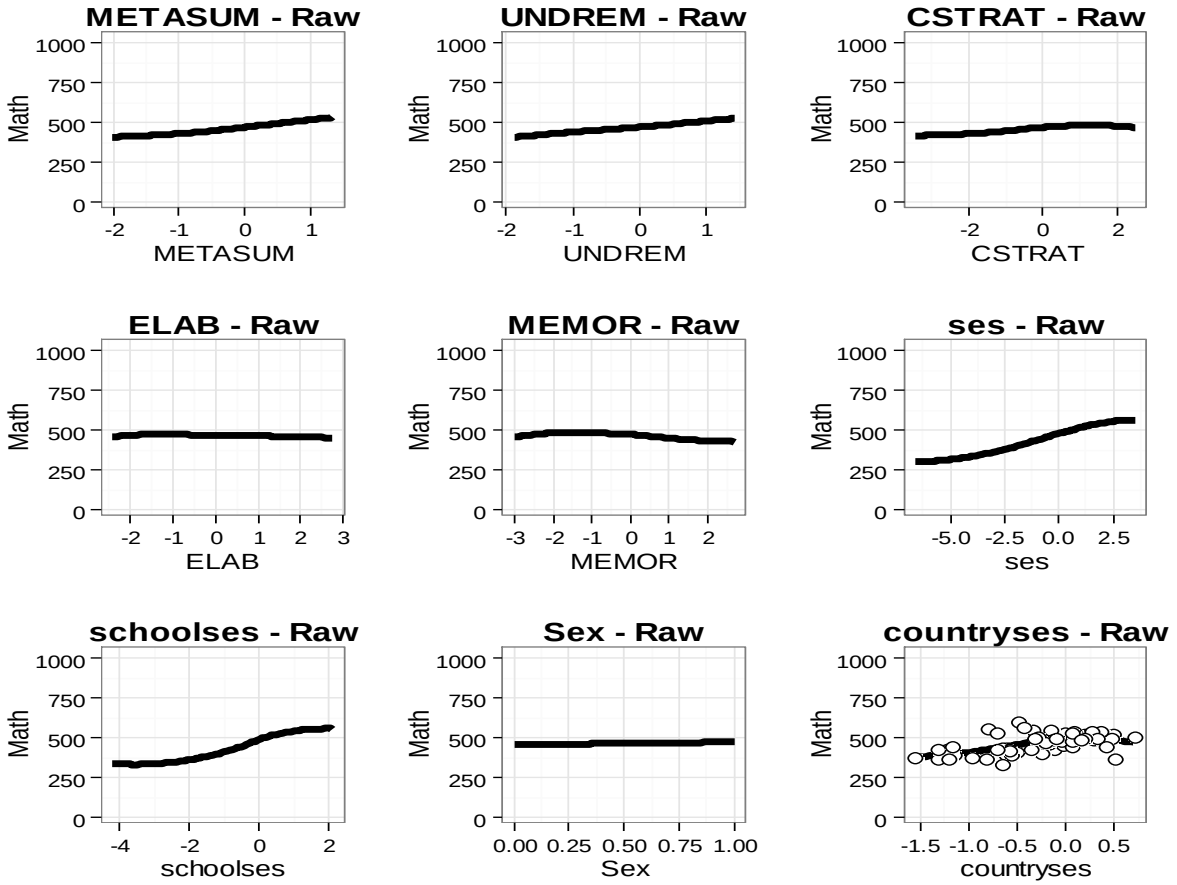

Figure 2. Predicted value plots for main effects of each predictor variable in the RF model.

relationship between SES and math achievement score appears to be stronger for individuals with lower METASUM scores. In other words, the relationship of SES on math test score is stronger for individuals who are less likely to use summarizing learning strategies. As a way to contrast the case where an interaction is present and where it is not, Figure 4 contains a set of interaction plots for SES and ELAB. In this case, it is clear that students with higher SES have higher math test s cores, that use of elaboration strategies does not appear to be related to performance on the test, and that there is no interaction between ELAB and SES with respect to math score, given the very similar shape of the lines for each of the three elaboration score groups in the figure. Similar graphs can be examined for other interactions of interest.

Table 1. Summary of RE-EM Tree Results

Terminal node	Node size	Math Mean
1	1138	393.4
2	689	410.6
3	735	441.9
4	493	418.5
5	772	448.9
6	330	455.3
7	785	489.0
8	458	443.9
9	392	474.5
10	258	466.3
11	464	504.3
12	364	494.5
13	521	530.2
14	364	494.5
15	521	530.2
16	1071	511.4
17	411	542.6
18	534	541.9
19	585	573.1

RE-EM Trees

Whereas the primary goal of MLEDA is understanding the extent to which individual variables and their interactions contribute to scores on the outcome variable, the major purpose of RE-EM trees is prediction. For this reason, the descriptive tools available to researchers using this technique are not as advanced as those for MLEDA. Nonetheless, RE-EM trees do provide useful information regarding relationships between the predictors and the outcome variable, including with respect to interactions among the predictors. Indeed, in certain respects the RE-EM tree approach yields more detailed information regarding these relationships than does MLEDA, as will be demonstrated below.

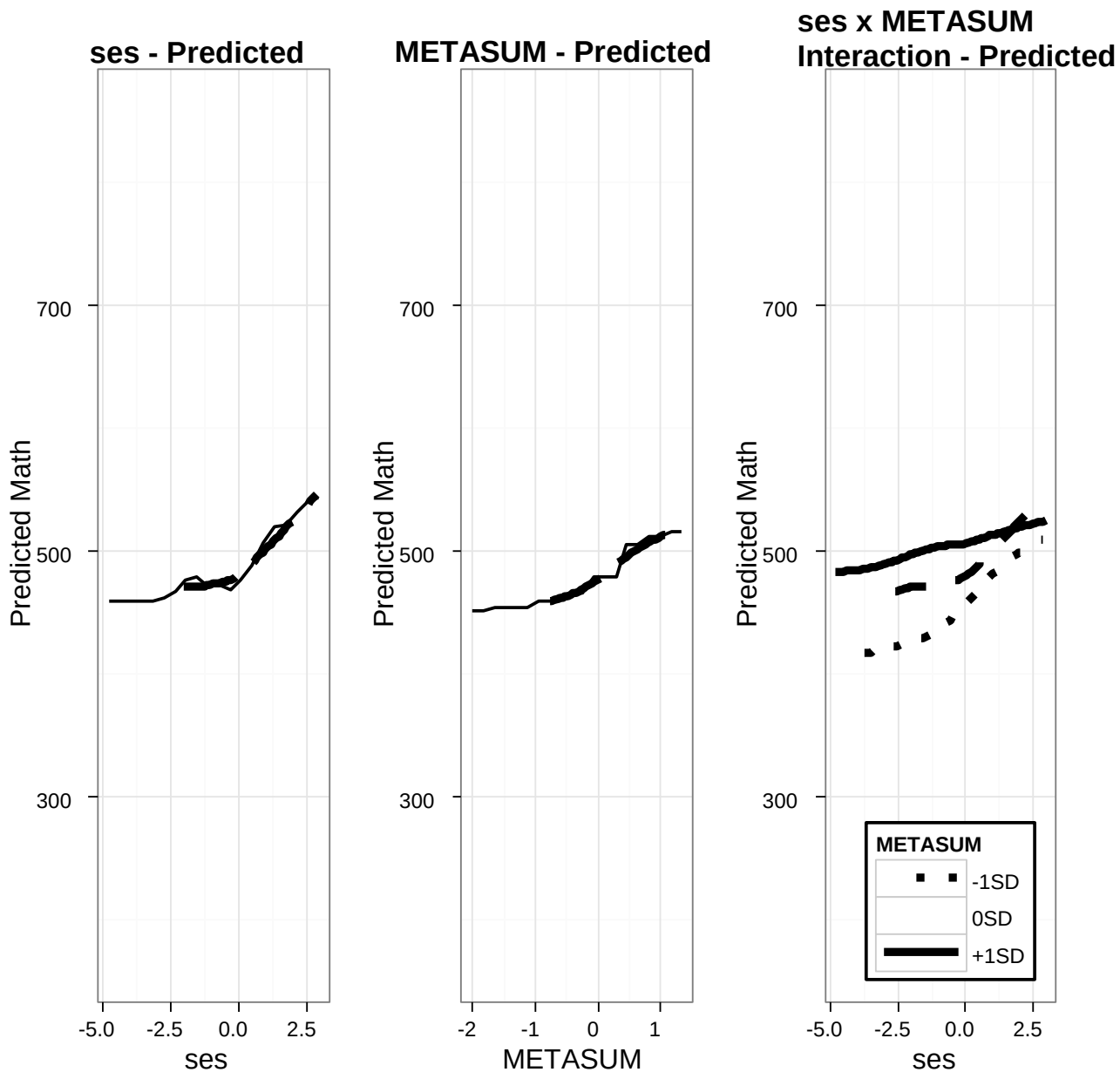

Figure 3. Predicted value plots for interaction of student SES, use of the summarizing learning strategy, and math achievement score in the RF model

Figure 5 displays the RE-EM tree with math achievement score as the dependent variable, and the predictors being those used with MLEDA, and described above. The first point to make regarding interpretation of the tree is that the length of the lines linking the nodes reflects the decrease in deviance (unexplained variance in *Y*) that is obtained through a particular split. Thus, it is clear that the first split reduces variance the most, followed by later splits in the tree. The final splits at the bottom decrease the unexplained variance in the outcome variable relatively little by comparison. This first split of the data was for school SES, such that individuals attending schools with mean standardized SES values less than -0.1596 (just slightly below average) were placed on the left side of the tree, and those with mean school SES values larger than -0.1596 were placed to the right. For individuals with lower school SES, the next split was on the summarizing score for metacognitive strategy use. Those with METASUM scores less than -0.1231 were moved to the left, and those with METASUM values larger than -0.1231 were moved to the right. It is possible to continue down the tree until the terminal nodes are reached. At the bottom of each terminal node RE-EM tree displays the mean value of the outcome variable, in this case math achievement score. Table 1 includes summary information for each of the terminal nodes, including node

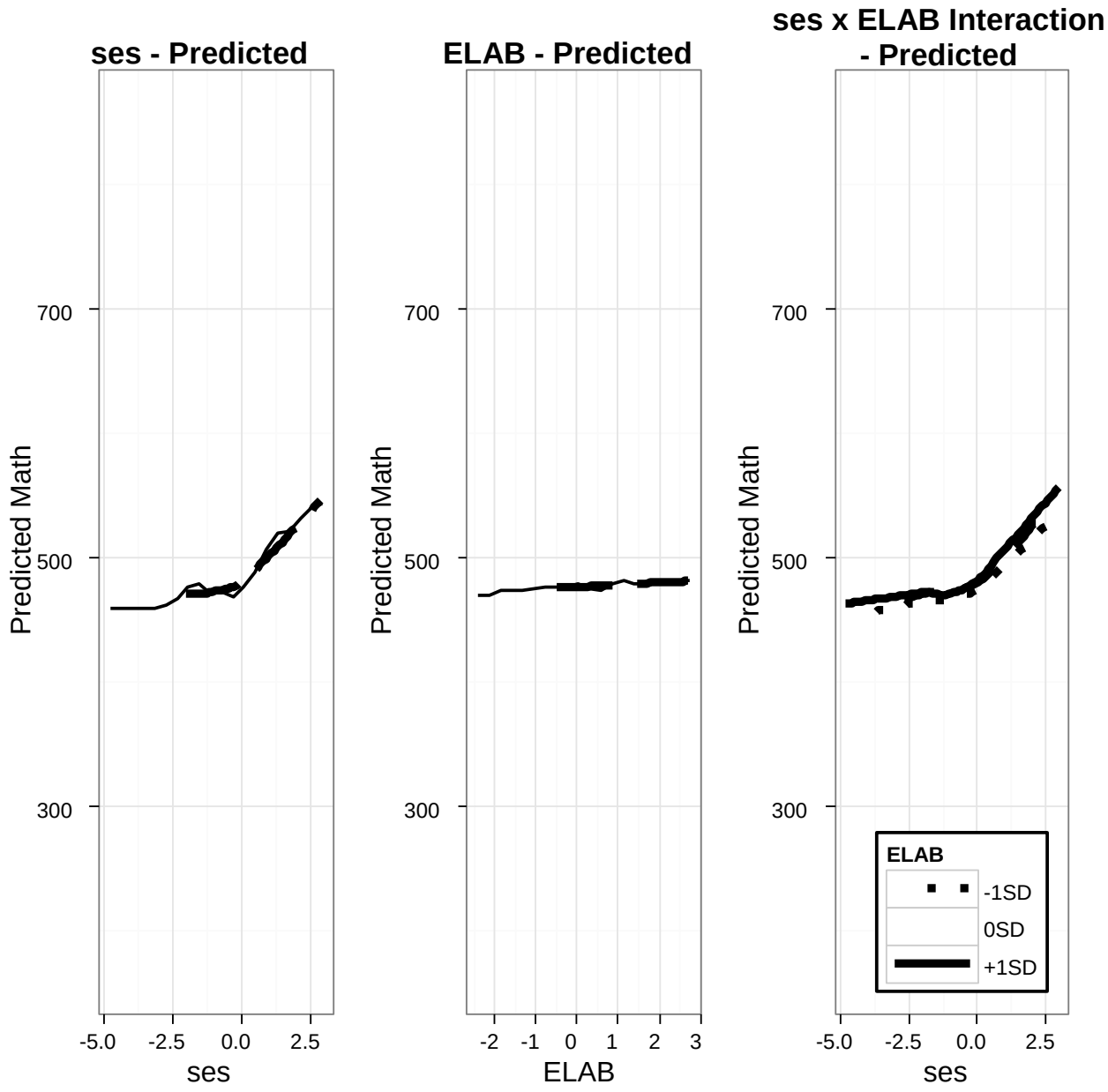

Figure 4. Predicted value plots for interaction of student SES, use of the elaboration learning strategy, and math achievement score in the RF model.

size and mean math score. From this table, we can see that terminal node 19 yielded the largest mean math achievement score. This node included individuals with school SES greater than 0.4073, METASUM greater than 0.2962, and UNDREM greater than 0.4992. In other words, these were individuals who attended relatively wealthy schools, and who used both summarizing and understanding/remembering learning strategies. In order to fully understand the relationships among the variables measured here and math achievement, a similar tracing of the paths down the tree can be done for each terminal node.

Discussion

The purpose of this study was to demonstrate two approaches for developing recursive partitioning models in the presence of multilevel data. These methods account for the data structure in different ways. MLEDA does not alter the basic algorithm underlying RF, but rather changes the way in which variable importance is calculated, by using a simulated cross-validation dataset rather than the typical out of bag sample that is used in standard RF analyses. In this way, the multilevel data structure is appropriately

accounted for and the artificial reduction in OOB error that occurs for standard RF is avoided, leading to more accurate measures of relative variable importance. In contrast, the RE-EM tree approach explicitly models random effects (e.g., intercepts and slopes), and uses residuals from this model in place of the actual dependent variable values, thereby accounting for the multilevel data structure. The two approaches differ in terms of their focus, with MLEDA being very much an exploratory tool providing researchers with global information regarding which predictors are most closely associated with the outcomes, and RE-EM tree displaying each specific split in the tree, along with sample sizes and means for the various terminal nodes.

As was noted previously, multilevel data presents researchers with methodological challenges, particularly with respect to correct identification of the importance of each predictor variable's relationship with the outcome. At the same time, given the ubiquity of such data structures in educational and psychological research, coupled with the need to use data mining techniques such as recursive partitioning with large, complex datasets, researchers need access to techniques that properly account for multilevel data structure. The two methods presented here do so, and provide the user with a variety of information about relationships in the data. In terms of which to use when, some recommendations can be made based upon the work presented here, as well as prior research in this area. First, if determining which independent variables and low level interactions of these are most strongly related to the outcome is of primary importance, then MLEDA may be preferable to RE-EM tree because it yields such information in an easy to interpret fashion. In addition, unlike RE-EM trees, MLEDA provides the researcher with an index of variable importance, such that the relative magnitude of relationships with *Y* can be compared. On the other hand, if the researcher would like to understand in more detail the specific mechanisms that differentiate individuals on the dependent variable with respect to the predictors, then RE-EM tree may be preferable. Whereas MLEDA yields its results in a black box way, such that the individual trees are not visible, Re-EM tree provides the user with the tree so that the paths to the terminal nodes can be clearly delineated. Thus, the mechanisms associated with appearance in one of the terminal nodes can see be clearly seen in the resulting trees. This type of result may be particularly interesting to those who are using RP to assist in development of a diagnostic framework for the outcome.

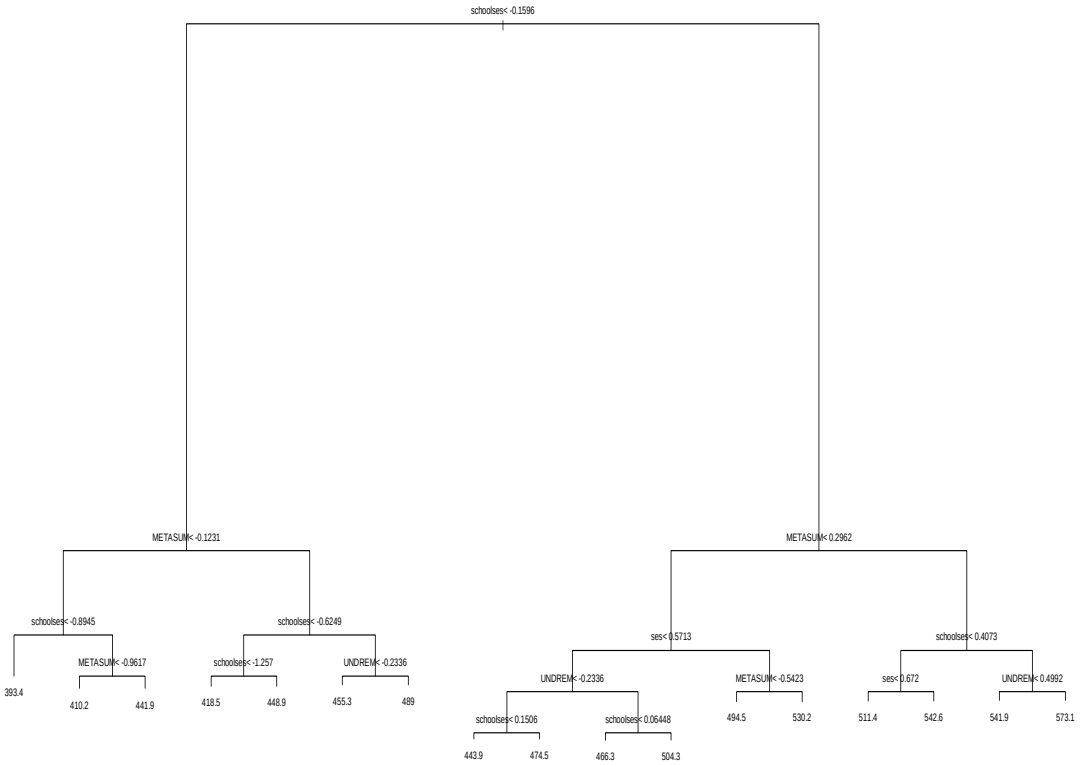

Figure 5. RE-EM tree for mathematics achievement with intercept random effect.

References

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning, 36,* 105-139.

Brieman, L. (1996). Bagging predictors. *Machine Learning, 24,* 123-140.

Brieman, L. (2001). Random forests. *Machine Learning, 45,* 5-32.

Brieman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth Publishing.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *Annals of Statistics, 30,* 927-961.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40,* 139-157.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests.* Boca Raton, FL: Chapman & Hall/CRC.

Finch, W. H., & Schneider, M. K. (2007). Classification accuracy of neural networks vs. discriminant analysis, logistic regression, and classification and regression trees: Three and five groups cases. *Methodology, 3*(2), 47-57.

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

Gruenewald, T. L., Mroczek, D. K., & Ryff, C. D. (2008). Diverse pathways to positive and negative affect in adulthood and later life: An integrative approach using recursive partitioning. *Developmental Psychology, 44*(2), 330-343.

Gumus, S. (2014). The effects of community factors on school participation in Turkey: A multilevel analysis. *International Review of Education, 50*(1), 79-98.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer-Verlag.

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement, 7*(5), 870-901.

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

Karpievitch, Y. V., Hill, E. G., Leclerc, A. P., Dabney, A. R., & Almeida, J. S. (2009). An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. *PloS One, 4*(9), 1-10.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage.

Lei, P., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *Journal of Experimental Education, 72,* 25-49.

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

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Markham, F., Young, M., & Doran, B. (2013). Detection of problem gambler subgroups using recursive partitioning. *International Journal of Mental Health and Addiction, 11*(3), 281-291.

Martin, D. P. (2015). *Efficiently exploring multilevel data with recursive partitioning*. (Doctoral dissertation). University of Virginia.

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

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behavior of random forest permutation-based variable importance measures under predictor correlation. *Bioinformatics, 11,* 110-122.

OECD. (2009). *Programme for International Student Assessment*.

Ozgen, H., Hellemann, G. S., & de Jonge, M. V. (2013). Predictive value of morphpologial features in patients with autism versus normal controls. *Journal of Autism and Developmental Disorders, 43*(1), 147-155.

Pai, D. R., Lawrence, K. D., Klimberg, R. K., & Lawrence, S. M. (2012). Analyzing the balancing of error rates for multi-group classification. *Expert Systems with Applications: An International Journal, 39*(17), 12869-12875.

Rausch, J. R., & Kelley, K. (2009). A comparison of linear and mixture models for discriminant analysis under nonnormality. *Behavior Research Methods, 41,* 85-98.

Ripley, B. D. (1996). *Pattern recognition and neural networks.* Cambridge, UK: Cambridge University Press.

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.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage Publications.

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

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

Van Laere, E., Aessaert, K., & van Braak, J. (2014). The role of students' home language in science achievement: A multilevel approach. *International Journal of Science Education, 36*(16), 2772-2794.

Williams, C. J., Lee, S. S., Fisher, R. A., & Dickerman, L. H. (1999). A comparison of statistical methods for prenatal screening for Down Syndrome. *Applied Stochastic Models in Business and Industry, 15*(2), 89-101.

Winnaar, L. D., Frempong, G., & Blignaut, R. (2015). Understanding school effects in South Africa using multilevel analysis: Findings from TIMMS 2011. *Electronic Journal of Research in Educational Psychology, 13*(1), 151-170.

Send correspondence to: W. Holmes Finch
Ball State University
Email: whfinch@bsu.edu

APPENDIX

```
attach(metacog.sample)

#Initial MLEDA plot
plot_ml(the_data = metacog,
        var_name = c("METASUM", "UNDREM", "CSTRAT", "ELAB", "MEMOR", "ses",
"schoolses", "Sex", "countryses"),
        var_level = c(1, 1, 1, 1, 1, 1, 1, 1, 2),
        cluster = "COUNTRY",
        outcome = "Math")

#Obtain the information needed to estimate ICC
null_mod <- lmer(Math ~ 1 + (1 | COUNTRY), data = metacog.sample)
summary(null_mod)

#Create cross validation samples
data_fold <- create_fold(metacog.sample, "COUNTRY", 2)


validate_ml(the_data = data_fold,
            formula = "Math ~ METASUM+ UNDREM+ CSTRAT+ ELAB+ MEMOR+
            ses+ schoolses+ Sex+ countryses",
            stat_method = "rf",
            cluster = "COUNTRY")

#Fit standard multilevel and random forest models using cross validation sample
naive_mod <- lmer(Math ~ METASUM+ UNDREM+ CSTRAT+ ELAB+ MEMOR+
            ses+ schoolses+ Sex+ countryses+(1|COUNTRY),
                  data = data_fold[data_fold$fold == 1, ])
```

```
rf_mod <- randomForest(Math ~ METASUM+ UNDREM+ CSTRAT+ ELAB+ MEMOR+
            ses+ schoolses+ Sex+ countryses,
                       data = data_fold[data_fold$fold == 1, ])

#Create list for variable importance
mod_list <- list(naive = naive_mod, cart_forest = rf_mod)

# Plot variable importance
importance_ml(mod_list)

#Plot School SES by SES
plot_ml(the_data = metacog.sample, the_mod = rf_mod, var_name = c("CSTRAT", "Sex"),
        var_level = c(1, 1), cluster = "COUNTRY", outcome = "Math", interact = TRUE)

plot_ml(the_data = metacog.sample, the_mod = rf_mod, var_name = c("ses",
"countryses"),
        var_level = c(1, 1), cluster = "COUNTRY", outcome = "Math", interact = TRUE)

plot_ml(the_data = metacog.sample, the_mod = rf_mod, var_name = c("Sex", "MEMOR"),
        var_level = c(1, 1), cluster = "COUNTRY", outcome = "Math", interact = TRUE)

plot_ml(the_data = metacog.sample, the_mod = rf_mod, var_name = c("ses",
"countryses", "schoolses"),
        var_level = c(1, 2, 1), cluster = "COUNTRY", outcome = "Math")

*****REEMtree*****
library(REEMtree)
attach(metacog.sample)

#Fit the random intercept REEM tree model
metacog.reemtree.intercept<-REEMtree(Math ~ METASUM+ UNDREM+ CSTRAT+ ELAB+ MEMOR+
            ses+ schoolses+ Sex+ countryses, data=metacog.sample, random=~1|COUNTRY,
tree.control(nobs, minsize=20))

#Plot the tree
plot(metacog.reemtree.intercept)

#Obtain summary information about the tree
summary(metacog.reemtree.intercept)
metacog.reemtree.intercept.best<-prune.tree(metacog.reemtree.intercept, best=10)

#Fit a random intercept and random slope for METASUM model
metacog.reemtree.slope.metacog<-REEMtree(Math ~ METASUM+ UNDREM+ CSTRAT+ ELAB+ MEMOR+
            ses+ schoolses+ Sex+ countryses, data=metacog.sample,
random=~1+METASUM|COUNTRY)

#Plot the tree
plot(metacog.reemtree.intercept)

#Obtain the summary information about the tree
summary(metacog.reemtree.slope.metacog)
```