Journal of Data Science 20 (3), 359–379
July 2022

DOI: 10.6339/22-JDS1056
Data Science in Action

Tree-Based Methods: A Tool for Modeling Nonlinear Complex Relationships and Generating New Insights from Data

Ya Mo[1,2,*], Brian Habing[2,3], and Nell Sedransk[2]

[1]*Department of Curriculum, Instruction, and Foundational Studies, College of Education, Boise State University, 1910 University Drive Boise, ID 83725-1745, U.S.A.*
[2]*National Institute of Statistical Sciences, Washington D.C., U.S.A.*
[3]*Department of Statistics, University of South Carolina, Columbia, U.S.A.*

Abstract

Our paper introduces tree-based methods, specifically classification and regression trees (CRT), to study student achievement. CRT allows data analysis to be driven by the data's internal structure. Thus, CRT can model complex nonlinear relationships and supplement traditional hypothesis-testing approaches to provide a fuller picture of the topic being studied. Using Early Childhood Longitudinal Study-Kindergarten 2011 data as a case study, our research investigated predictors from students' demographic backgrounds to ascertain their relationships to students' academic performance and achievement gains in reading and math. In our study, CRT displays complex patterns between predictors and outcomes; more specifically, the patterns illuminated by the regression trees differ across the subject areas (i.e., reading and math) and between the performance levels and achievement gains. Through the use of real-world assessment datasets, this article demonstrates the strengths and limitations of CRT when analyzing student achievement data as well as the challenges. When achievement data such as achievement gains in our case study are not linearly strongly related to any continuous predictors, regression trees may make more accurate predictions than general linear models and produce results that are easier to interpret. Our study illustrates scenarios when CRT on achievement data is most appropriate and beneficial.

Keywords *achievement; early childhood education; tree-based methods*

1 Introduction

1.1 Background

Linear models have been widely used in educational research because they often show adequate fit, and perhaps because they are the most commonly taught (Yan and Su, 2009). However, there are also times that alternative nonlinear models have not been fully explored; as a result, a nonlinear complex relationship may not be uncovered, and opportunities for new insights from data are missed. In comparing the use of three types of nonlinear models and linear regression modeling on school-level data on 183 elementary schools, Baker (2001) found that flexible modeling raised unique questions and identified nonlinear relationships that linear regression modeling overlooked.

*Corresponding author. Email: yamo@boisestate.edu.

Received January 1, 2022; Accepted June 21, 2022

Baker's observation carries contemporary importance because one consequence of the diversity of the US is that differences in relationships between predictors and outcomes across the subgroups could be a common occurrence. Further, reasonably homogeneous subgroups may not be definable by just a single student characteristic, such as student's ethnicity, socioeconomic status, gender, or special status as English Language Learners (ELLs) or Individual Educational Plan (IEP) status. Instead, relevant subgroups are likely to be defined by a combination student and/or school characteristics. For example, the relationships between predictors and outcomes may differ for students of different ethnicities from families with higher or lower socioeconomic status and may also depend on whether the student is classified as ELL or IEP.

This kind of nonlinearity can be illustrated by posing questions. When boys' and girls' performances differ, is that observed difference of the same magnitude and in the same direction at the lower and higher ends as well as the middle of a socioeconomic status scale? Is that same difference seen in every ethnic group? Is the same difference observed in ELLs and IEPs in different ethnic groups at each socioeconomic level?

Although there is a rich class of linear models, and they can include interactions among predictors, each term in the model must apply to the whole population unless a term of one order higher was used to separate subgroups. Terms including the interactions have to be pre-specified if automatic model building methods are not used, and the higher-order interaction terms are often hard to interpret. Even without interactions, multicollinearity can complicate interpretation. A nonlinear model that does not require the pre-specification of the interactions among predictors can significantly aid in understanding the educational phenomenon being studied and in generating new insights from data.

1.2 Purpose

This manuscript aims to introduce tree-based methods, in particular, classification and regression trees (CRT), to study student achievement. CRT can be used (1) directly to allow data to be driven by its internal structure, (2) to supplement existing hypotheses testing to provide a fuller picture of the topic being studied, and (3) in combination with other modeling approaches to allow a hybrid modeling approach. Section 2 of this paper reviews the concepts of recursive partitioning and regression tree algorithms. Sections 3 and 4 focus on the regression tree for uses (1) and (2) for student achievement. The research on the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011) provides a case study to demonstrate the application of the regression tree. In particular, we used kindergarten year fall term and spring term assessments of performance in reading and in math to investigate the impact of factors from students' demographic backgrounds on those reading and math skill levels. The analysis of students' gains in math and reading skills drew on both fall term and spring term assessment data from 2011 ECLS-K. A similar approach using a nonlinear analysis yielded findings of important differences among the subpopulations identified. Section 5 briefly discusses additional uses of tree-based methods in educational research and the use of classification and regression tree in possible hybrid modeling approaches. Through the use of real-world assessment datasets, this article will demonstrate the strengths and limitations of CRT when analyzing student achievement data as well as the challenges. Such illustration will contribute to the use of CRT on achievement data by highlighting scenarios it is most appropriate and beneficial.

2 Overview of Classification and Regression Tree Algorithms

2.1 Concept for Recursive Partitioning

CRT is a simple-to-use but powerful tool for analyzing complex data that can be used for exploration, description, and prediction. CRT and other recursive partitioning algorithms are widely used in various fields of study (e.g., ecology, medicine, genetics, bioinformatics), but have only relatively recently been introduced into educational research and applied in areas such as studying the rate of growth in educational achievement and drop-out and educational measurement modeling (e.g., Rupp et al., 2001; Ma, 2005; Jeon and De Boeck, 2016; Jeon et al., 2017).

Recursive partitioning algorithms including CRT are a class of machine learning techniques that partition a population into a set of internally homogeneous and mutually exclusive subgroups, where the defining factors need not be the same for all subgroups (Ma, 2018). Inference can then be made about the role of factors at both subgroup and population level; and predictions can be made and tested based on the subgroup profiles.

The recursive partitioning algorithms create generations of subsets. Starting with the whole population (i.e., the first parent node), child nodes are created by selecting one "best factor" to use in splitting the data into two or more subsets and by determining the optimal cut point(s) for that factor. These child nodes become the next generation of parent nodes. The key is that in this next generation each new parent node is individually split using the optimal choice of factor and the cut points for that node only. A chain of nested subsets terminates, and the process ends (separately for each subset chain) when a later-generation child node cannot be profitably split. The tree that is created is commonly annotated with information about the split and the resulting subsets. Depending on the outcome variable and the purpose of the analysis, CRT algorithms have a variety of possible specifications. For example, a binary outcome variable leads to a classification algorithm and the criterion for predictor selection is usually some measure of misclassification. In contrast, regression may be the basis when the outcome variable is continuous and the criterion may be regression sufficiency (e.g., R^2, F-statistic or other goodness-of-fit measure) or regression coefficient precision (e.g., CV or p-value). Typical termination rules include minimum values for the criterion and/or minimum sample size in the terminal nodes.

2.2 Comparison of Trees and Linear Models

For the case of continuous response variables, the data sets where CRT may be applicable can also be approached using general linear models such as regression, ANOVA, or ANCOVA. When the relationship between outcome variables and predictors are essentially linear, linear models are likely to outperform the trees. However, when the relationship between outcome variables and predictors are nonlinear and complex, tree-based methods are likely to outperform linear models in prediction accuracy. Furthermore, in such cases the tree-based methods are preferable because of their interpretability and visualization (James et al., 2017).

Similarity CRT faces a similar challenge to that of linear models when using stepwise procedures instead of "best subsets" procedures – the procedure for selecting the best partition is optimal at each step but not overall. When stepwise variable selection is used, variables are selected considering what was entered before and not considering what variables yet to come. More specifically, among all the possible partitions, the partition chosen is optimal for the particular

step of the tree growing instead of being optimal for the entire tree (Breiman et al., 1984); this type of search method is called a "greedy" search. The greedy algorithms have the weakness of sometimes not finding globally optimal solution by not considering all the possible solutions.

Difference While the different predictors are combined linearly in linear regression, CRT consists of rectangular partitions derived from recursive splitting, including nonlinear and non-monotone association rules (Strobl et al., 2009). Further, CRT does not assume a normal error distribution; thus, making it easier for researchers in terms of not having to deal with the consequences of violating distributional assumptions. The interactions included in the linear regression are typically limited to the two-way interactions specified before the fitting process; on the other hand, "only the interactions that are actually used in the tree are generated during the fitting process" (Strobl et al., 2009, p. 5). In addition, categorical variables may need to be dummy coded in linear models, but CRT can be applied to any data structure with various kinds of predictors (e.g., sparse, skewed), including both categorical and continuous data without a necessity of pre-processing them (Kuhn and Johnson, 2013).

Tree-based methods resemble human's decision-making process more than traditional regression and classification approaches (James et al., 2017). When reporting the results, CRT presents the results in an easy-to-interpret form; the tree classifiers illuminate the structure of the data and the steps one should use to decide. However, less-than-optimal predictive performance can happen when the relationship between the independent variables and the outcome variable cannot be best captured by the rectangular homogenous subspaces defined by the tree partitions. Ensemble methods that combine multiple trees and average over them, such as random forests, have shown great improvement in the prediction accuracy, although with a loss in interpretability (James et al., 2017; Kuhn and Johnson, 2013; Strobl et al., 2009).

2.3 CHAID Algorithm

The CHAID (Chi-square automatic interaction detection; Kass, 1980) is a stepwise CRT procedure that uses chi-square statistics (for categorical outcomes) and F-statistics (for continuous outcomes) to measure the relationship between the outcome and categorical or continuous predictor variables. Instead of binary splits of the predictor variables, CHAID would ideally search through all possible groupings of a categorical predictor or groupings of a continuous predictor (Ledolter, 2013). In order to avoid the time-consuming process of searching through all possible groupings, a three step simplification is used: (1) If there is a continuous predictor, the predictor is divided into categories with an approximately equal number of cases in each category; (2) For a particular node and a particular predictor, if there is not a significant association between a pair of the categories of the predictor and the outcome variable at a predefined alpha-to-merge value, it will merge the pair of the predictor categories and iterate this step; if there is a significant association, a Bonferroni adjusted p-value will be computed for the set of categories; and (3) The predictor variable with the smallest Bonferroni adjusted p-value will be selected as the split variable if its adjusted p-value is smaller than a predefined alpha-to-split value; otherwise, there will be no further splits; the nodes become terminal nodes (Ma, 2018; Ledolter, 2013). An advantage of CHAID among CRT methods is that it assigns more or less the same number of cases to each node, thus not creating extreme terminal nodes, which allows for drawing appropriate policy and practice implications (Ma, 2018).

The default alpha-to-split value and alpha-to-merge values for the CHAID algorithm in SPSS Modeler 18.3 (IBM Corp, 2021a) are both .05. For classification problems, Pearson Chi-

square statistics are used as the default for the Chi-square test with the likelihood-ratio Chi-square is an alternative; for regression problems, F-tests are used. For the estimation of the CHAID model, the default convergence value is .05, and the default maximum number of iterations is 100. The Bonferroni correction is the default of adjusting for significance values with multiple comparisons. A continuous predictor variable is banded into ten discrete intervals by default, but the number of discrete intervals can be increased up to 64 for and will be applied to all scale predictors in the regression tree.

In the following Sections 3 and 4, a case study using the 2011 ECLS-K data will demonstrate the application of CRT using the CHAID algorithm on a continuous response, and will compare the regression tree results to linear model results.

3 Case Study: Application of Regression Tree on ECLS-K Study

3.1 Background

With the increasing racial and ethnic diversity in the US student population and the increasing economic disparities between advantaged and disadvantaged students, the student population continues to undergo change. The present study focuses on a few key demographic variables that are routinely reported as being related to math and reading performance: students' gender, ethnicity, SES, schools' characteristics including school percent free or reduced-price lunch (% FRPL) and geographic designation (Mulligan et al., 2012). This study contributes to the pool of research examining the complex and the intricate relationships among students' and schools' demographic factors and students' math and reading performance in early childhood education (e.g., Cheadle, 2008; Cooper et al., 2010). Differing from previous studies, this study utilized a data-driven analytic approach – regression tree analyses – to display the patterns of the interplay between demographic factors in relation to students' outcomes. This study highlights the strength of using regression tree on achievement data, such as its interpretability and utilizing best predictors for subgroups, but also discusses the challenges as student achievement is often strongly linearly related to dominant continuous predictors such as students' prior achievement or students' socioeconomic (SES) status.

3.2 Data and Sample

The ECLS-K Fall 2010 and Spring 2011 data were used for this study. ECLS-K: 2011 provides information on children's early school experiences by following a nationally representative sample of children from kindergarten through elementary school. For kindergarten students, it consists of one-on-one assessments of children, interviews with parents, and self-administered questionnaires from teachers, school administrators, and nonparental out-of-school care providers.

A total of approximately 18,170 kindergartners from about 1,310 schools participated in the ECLS-K study. Their parents, teachers, school administrators, and out-of-school care providers also supplied information about them through participating in the interviews and answering questionnaires (Tourangeau et al., 2015). Approximately 12,900 kindergarten students who have valid math and reading achievement scores were selected for the sample. Around 850 kindergarten students were excluded because of missing information or because their race was not ascertained, or numbers of students in the ethnic group were too small for analysis.

3.3 Variables

Outcomes Students' item response theory (IRT)-based scale scores are used as outcome variables because (1) they are the overall measures of students' achievements, (2) can be compared for both cross-sectional and longitudinal analyses, (3) can be compared among subgroups of children, (4) can be used to study the correlations between achievement and children's background, and (5) can be used to calculate a gain score by taking the difference between spring and fall scores (Tourangeau et al., 2015). Students' fall performance allows us to investigate the important factors that are closely related to students' achievement prior to their formal schooling experience. Students' spring performance allows us to investigate the relationship between students' academic achievement and the same predictors while controlling for students' prior achievement (i.e., fall performance). This approach, as well as a direct approach using the gain scores as the outcome variables, allows us to examine students' achievement gains between the fall performance and spring performance in reading and math. More specifically, the following outcomes are used:

- Fall Performance Level: 2011 reading & math results
- Spring Performance Level: 2011 reading & math results
- Achievement Gain: Difference in scores (Spring 2011 score – Fall 2010 score) in reading & math for each individual student

Predictors The following predictors are used in the regression tree analysis:

- Socioeconomic Status (SES) – individual student level index value
- Race – White (non-Hispanic), Black/African American (non-Hispanic), Asian (non-Hispanic), Hispanic (subgroups: race specified/ not specified)
- Location type – Urban (subgrouped by size), Suburban (subgrouped by size), Town (subgroup by remoteness), Rural (subgroup by degree), which was recoded into five categories: rural, town, suburban, small and middle city, and large city.
- Gender
- School percent free or reduced-price lunch (% FRPL)
- Region – Northeast, South, Midwest, West
- Fall Performance Level: 2011 reading & math results (used as predictors when spring performance levels are outcome variables).

Descriptive statistics of these variables can be found in Table 1 and Table 2.

3.4 Statistical Analyses

Regression tree analyses were conducted using the CHAID algorithm separately for each of the six outcome variables: fall performance in reading, fall performance in math, spring performance in reading, spring performance in math, kindergarten gain in reading, kindergarten gain in math. CHAID is used because, unlike other CRT methods, it allows for multi-way node splitting – allowing for more complex decision rules with the many nominal and continuous scale predictor variables. Because of this, it has the advantage of displaying the complex relationship among predictors. Although not used in this study, the patterns or classification captured by CHAID can be further utilized in conjunction with other statistical models.

Five-fold cross-validation was used to evaluate the performance of the regression trees; the whole dataset was randomly divided into 5 subsets, and each time, one of the subsets served as the testing set, while the other four subsets together served as the training set the model was

Table 1: Frequency table of categorical variables.

Variable	Category	N	%
Location Type (Fall)			
	City, Large	2030	15.7
	City, Midsize & Small	2220	17.2
	Suburb	4450	34.5
	Town	1070	8.3
	Rural	3120	24.2
	Missing	10	
Location Type (Spring)			
	City, Large	2030	15.7
	City, Midsize & Small	2230	17.3
	Suburb	4450	34.5
	Town	1070	8.3
	Rural	3120	24.2
Region (Fall)			
	Northeast	1910	14.8
	Midwest	2840	22.0
	South	4840	37.5
	West	3320	25.7
Region (Spring)			
	Northeast	1910	14.8
	Midwest	2830	22.0
	South	4830	37.5
	West	3320	25.7
Race			
	White	6300	48.8
	Black	1860	14.4
	Hispanic	3650	28.3
	Asian	1090	8.4
Gender			
	Male	6620	51.3
	Female	6280	48.7

NOTE: Due to data security restrictions, unweighted sample size numbers are rounded to the nearest ten. N = Sample Size; % = Percentage.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K), Fall 2010 and Spring 2011.

fit to. The mean absolute error was calculated as the deviation of the training sample derived predicted value from the observed value in the testing set, and was recorded and averaged across the five sets. This was done for trees containing from one to six levels, and the mean absolute error was compared across trees of different depths. Having the noticeably smallest mean absolute error with the lowest tree depth led a tree to be selected as the final regression tree for an outcome variable. As a result, 2 levels was selected for the fall math performance and spring reading performance; 3 levels was selected for the spring math performance and reading

Table 2: Descriptive statistics of continuous variables.

	N	Range	Minimum	Maximum	Mean	Standard Deviation	Variance
School %FRPL	12900	100.00	.00	100.00	48.6733	30.07002	904.206
SES	11620	4.93	−2.33	2.60	−.1381	.79854	.638
Reading Gain	12040	112.48	−57.06	55.42	14.2692	7.94190	63.074
Math Gain	11970	66.31	−22.83	43.48	13.5543	7.12757	50.802
Fall Reading Performance	12200	84.47	25.45	109.92	46.5265	11.44929	131.086
Spring Reading Performance	12480	84.24	25.68	109.92	60.6588	13.33363	177.786
Fall Math Performance	12150	104.39	7.19	111.58	31.0955	11.29931	127.675
Spring Math Performance	12450	81.57	7.19	88.76	44.5003	12.18949	148.584

NOTE: Due to data security restrictions, unweighted sample size numbers are rounded to the nearest ten.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K), Fall 2010 and Spring 2011.

gain. Though a tree of five level depth had the least mean absolute error for the fall reading performance because the mean absolute error between the three-level tree and the five level tree was only 0.004, while the two extra levels made the tree much harder to interpret, a three level tree was retained for the fall reading performance as the final regression tree. The same rationale was applied to select a two-level tree as the final regression tree for math gain. Please see Table 3 for the mean absolute error comparison.

3.5 Results

The relative importance of each predictor in relation to the outcome variable differed between fall performance, spring performance and gain during kindergarten; there were also important differences in predictors' relative importance between reading and math.

Analysis of Math Fall Performance SES, the individual student socioeconomic index value, is the predictor with the strongest (positively correlated) relationship with students' fall 2011 math scores (See Figure 1 for the tree structure). The algorithm divided the range for SES into 10 segments, with the mean scores for students in these 10 groups increasing by approximately 1.860 points (ranging from 1.161 to 2.857) from lower to next level SES group.

For each of the first 7 (lowest) socioeconomic subgroupings that include 69.95% of the kindergarteners as well as the upper subgrouping that contains 19.90% of kindergarteners on the SES scale, the pattern within the subgroup was essentially the same. Within a socioeconomically similar group, racial group was the strongest indicator of math score. While the algorithm sometimes broke the subgroup into two or three further subgroups, these followed a similar pattern: Black students and Hispanic students with lowest (mean) score; Asian students with highest (mean) score.

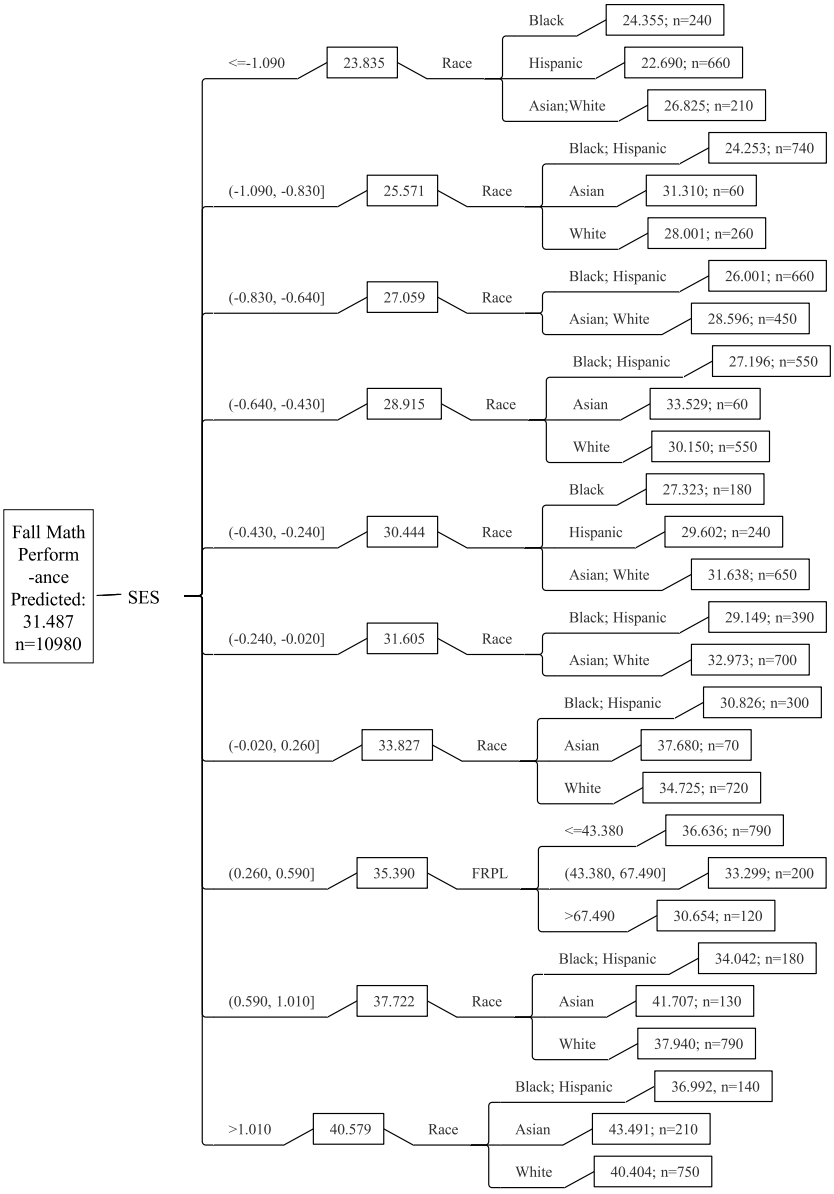
Table 3: Mean absolute error with cross-validation.

	ANCOVA		**ANCOVA with Interactions**		**Tree (1 level)**		**Tree (2 level)**	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Fall Reading Performance	7.661	7.680	7.616	7.650	7.792	7.809	7.696	7.773
Fall Math Performance	7.875	7.893	7.835	7.880	8.029	8.043	7.893	**7.952**
Spring Reading Performance	6.075	6.086	6.000	6.050	6.398	6.395	6.326	**6.371**
Spring Math Performance	5.494	5.500	5.429	5.456	5.712	5.718	5.653	5.704
Reading Gain	6.422	6.439	6.364	6.421	6.172	6.183	6.137	6.156
Math Gain	5.944	5.958	5.915	5.952	5.655	5.658	5.634	5.641
	Tree (3 level)		**Tree (4 level)**		**Tree (5 level)**		**Tree (6 level)**	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing
Fall Reading Performance	7.654	7.759	7.643	7.758	7.640	**7.755**	7.640	7.755
Fall Math Performance	7.851	7.960	7.835	7.971	7.831	7.970	7.831	7.970
Spring Reading Performance	6.285	6.377	6.273	6.381	6.271	6.380	6.271	6.380
Spring Math Performance	5.623	**5.699**	5.613	5.703	5.612	5.704	5.612	5.704
Reading Gain	6.114	**6.146**	6.080	6.148	6.066	6.152	6.061	6.154
Math Gain	5.611	5.643	5.589	**5.628**	5.582	5.631	5.580	5.632

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11(ECLS-K), Fall 2010 and Spring 2011.

Differences between highest and lowest racial group mean scores within socioeconomic subgroup averaged about 5.108 points (ranging from 2.595 to 7.665). White students sometimes grouped with Asian students and sometimes formed a middle group on their own. Black/African American students and Hispanic students often grouped together and scored below Whites. For 10.15% of kindergarteners (the next to two highest SES groups) on the SES scale, the pattern is slightly different and the school economic index becomes important as the factor creating the second subdivision. For the two next-to-highest SES groups the final subsetting was according to the school economic index, with scores inversely related to % FRPL.

Analysis of Math Spring Performance Math fall performance is the predictor with the strongest (positively correlated) relationship with students' spring 2011 math scores (See Figure 2 in the supplementary material for the tree structure). The algorithm divided the range for math fall performance into 10 segments, with the mean scores for students in these 10 groups increasing by approximately 3.912 points (ranging from 2.466 to 7.561) from lower to next level fall performance group.



NOTE: Due to data security restrictions, unweighted sample size numbers are rounded to the nearest ten.
SES=Socioeconomic Status; FRPL=% Free or Reduced-Price Lunch.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K), Fall 2010 and Spring 2011

Figure 1: Regression tree of fall math performance.

For each of the first (lowest) 4 fall performance subgroupings that include 40.00% of the kindergarteners and for each of the upper middle 3 fall performance subgroups that include 30.00% of the kindergarteners, the pattern within the subgroup was essentially the same. Within a fall performance similar group, racial group was the strongest indicator of math score. While the algorithm sometimes broke the subgroup into two and sometimes into three further subgroups, these followed a similar pattern with two exceptions: Black students with lowest (mean) score, often grouped with Hispanic students, and in a rare case with White students in the lowest

fall performance group; Asian students with highest (mean) score, often grouped with White students, and in a rare case grouped with Hispanic students in the middle fall performance group (33.489, 36.458].

For the upper 20.00% of kindergarteners on the fall performance scale, the pattern is slightly different and SES becomes important either as the factor creating the second subdivision (the highest fall performance group) or the third subdivision of the next to highest fall performance group with students from Midwest. The CHAID algorithm subsets the next to highest SES group according to region. The differences in mean scores by region are consistent with the Midwest mean score often being the highest; mean scores for West and South either fall between or are grouped with Northeast. In a rare case, for Black and Hispanic students with the upper middle fall performance (36.458, 41.034], those from West scored approximately 2.407 points higher than those from Midwest, South, and Northeast grouped together.

For the middle fall performance (27.619, 30.619] group, the algorithm created subsets with location type with rural students' mean scores approximately 1.968 points higher than those from other location types. Below the level of the second subdivision, the algorithm created further subsets with the SES and race group. For rural students in the middle fall performance group, race made a difference; for students from other location types in the same fall performance group, SES mattered. Gender entered the regression tree as a third subdivision in the highest fall performance group (8.77% students) and in the middle fall performance group (6.13% students), with female mean score 2.108 points higher than male mean score.

Analysis of Math Gain (Spring 2011 Performance – Fall 2010 Performance) Unlike spring performance, for Gain in Math the strongest relationship is to racial group (See Figure 3 in the supplementary material for the tree structure). Hispanic and White kindergarteners demonstrate a greater gain in math performance level between Fall 2010 and Spring 2011 (13.773 point gain) than Asian and Black/African American grouped together (12.441 point gain). Among White and Hispanic kindergarteners, the next most highly related factor is region; for both groups Midwest and West mean gain is highest and Northeast is lowest (1.758 point difference) with South in between. These regional differences follow the same pattern as is seen at a lower level relative importance for Spring 2011 Math performance.

For Whites and Hispanics in the Midwest, West, and South, gains in math during the kindergarten year are related to school location (urban to rural). However, in the Northeast, an economic factor (i.e., school % FRPL) is most strongly related to gain in math. For Asian and Black/African American kindergartners, the CHAID algorithm does not subset by region at any point. Rather, this group is subdivided based on an economic factor (% FRPL) and then within each % FRPL group separated into the two locale type groups that differ in gain about 1.7 points (students from cities and town gaining more than those from suburb and rural areas in the $\leqslant$ 67.49% FRPL group; students from city, suburb, and rural areas gaining more than those from towns in the >67.49% FRPL group).

Analysis of Reading Fall Performance Similar to Math fall performance level, SES is the predictor with the strongest (positively correlated) relationship with students' fall 2011 reading scores (See Figure 4 in the supplementary material for the tree structure). The algorithm partitioned the range for SES into 9 segments, with the mean scores for students in these 9 groups increasing by approximately 1.977 points (ranging from 1.261 to 2.963) from each lower to the next level SES group.

For each of the first (lowest) 2 socioeconomic subgroupings and the last (highest) 2 socioeconomic subgroupings together comprising 39.73% of the kindergarteners, the pattern within the subgroup was essentially the same and similar to those patterns in math. Within each of these socioeconomically similar groups, racial group was the strongest indicator of reading score. Differences between highest and lowest racial group mean scores within socioeconomic subgroup averaged about 6.152 points (ranging from 2.702 to 10.215). Race was also important creating the third subdivision of the middle SES subgrouping for different regions with an averaged difference between highest and lowest racial group mean scores of 3.363 points (ranging from 2.541 and 4.185).

For the middle 30.62% of the kindergarteners on the SES scale, region was the most important factor creating the second subdivision, affecting most subgroups; students from South had highest mean scores, approximately 3.000 points (ranging from 2.339 to 3.647) higher than the lowest region group mean scores. Region was also important creating the third subdivision of Blacks and Whites in the top SES group and Hispanics in the bottom SES group as well as for both males and females in the middle SES group.

Besides the above subdivision, below the level of the second subdivision, the algorithm created further subsets for other SES groups, gender (rather than racial group) was a predominating factor. The differences in mean scores by gender are consistent with the female mean score approximately 2 points above male mean score. Gender was also important creating the third subdivision of Asians and Whites in the next-to-the-bottom SES group and the Midwest, Northeast, and South in the middle of the SES scale. School level % FRPL affected the middle 6.02% of the kindergarteners on the SES scale.

Analysis of Reading Spring Performance Similar to Math spring performance level, fall performance score is the predictor with the strongest (positively correlated) relationship with students' spring 2011 reading scores (See Figure 5 in the supplementary material for the tree structure). The algorithm partitioned the range for fall performance scores into 10 segments, with the mean scores for students in these 10 groups increasing by approximately 4.263 points (ranging from 2.116 to 12.080) from each lower to the next level fall performance group.

For each of the first next to lowest 4 fall performance subgroupings and the middle one subgrouping together comprising 50.00% of the kindergarteners, the pattern within the subgroup was essentially the same and similar to those patterns in math. Within each of these fall performance similar groups, racial group was the strongest indicator of reading score. Differences between highest and lowest racial group mean scores within fall performance subgroup averaged about 2.633 points (ranging from 1.908 to 4.067).

For the top 20.00% of the kindergarteners on the fall performance scale and 10.00% of the kindergarteners in the middle of the fall performance scale, SES (rather than racial group) was the most important factor creating the second subdivision. For the next to highest 10.00% of the kindergarteners on the fall performance scale and the 10.00% of the kindergarteners in the lowest fall performance group, gender was the most important factor creating the second subdivision. The differences in mean scores by gender are consistent with the female mean score approximately 1.734 points above male mean score.

Analysis of Reading Gain (Spring 2011 Performance – Fall 2010 Performance) The gain in reading was most strongly related to location type (See Figure 6 in the supplementary material for the tree structure). Students in the rural area and town area have the greatest gain

in reading (15.151 points), followed by students in the suburb area, and midsize and small size city (13.995 points), and then by students in large city (13.427 points).

In rural area and town, gain in reading is related to the school economic index (% FRPL). The gain is greatest in schools with 34.44–43.38% FRPL (16.776 points); followed by schools with 18.03–34.44% FRPL (15.747 points), schools with above 43.38% FRPL (14.894 points), schools with 8.09–18.03% FRPL (14.470 points), and lowest in schools with below 8.09% FRPL (12.342 points).

In city and suburb, gain in reading is related to race. In midsize and small city as well as suburb, White students had the greatest gain (14.537 points), followed by Hispanic and Asian (13.833 points), and then by Black (12.835 points). In large city, White and Asian students (14.676 points) showed greater gain than Hispanic and Black students (12.740 points). Below the level of the second subdivision, the algorithm created further subsets, gender, region, and school % FRPL were predominating factors.

4 Case Study: Comparison of Regression Tree and Linear Regression Model Results

4.1 Added Value of Non-Linear Model

General linear models (Analyses of Covariance (ANCOVA)) were fit corresponding to each of the six regression trees presented earlier to illustrate the different insights that can be gained from linear models by comparing them with regression tree results. The ANCOVA were conducted using SPSS 28 (IBM Corp, 2021b).

To investigate the interaction of predictors, a two-way interaction was added to the ANCOVA model for each outcome pair one at a time; then all the significant two-way interactions were retained in the ANCOVA model; if an interaction term became no longer significant in the presence of other interaction terms, the interaction term was dropped in the order from the least statistically significant to statistically not significant until all the interaction terms in the ANCOVA model were significant. The mean absolute error of ANCOVA with all the significant interaction terms were compared with those of ANCOVA with individual predictors only, the difference was in a range of 0.012 to 0.037 for each outcome variable; similarly, the difference of adjusted R^2 was in a range of 0.005–0.014 for each outcome variable. Thus, the ANCOVA model with individual predictors was retained as the final ANCOVA model for each outcome variable because of its interpretability and was compared with the regression tree results.

Assumptions for the linear models, including linearity, and normality and homoscedasticity of errors, were checked using plots. The corresponding P-P plots for the residuals were checked for the outcomes: Math Fall Performance Level, Math Spring Performance Level, Math Gain, Reading Fall Performance Level, Reading Spring Performance Level, and Reading Gain. The P-P plots showed that the distributions appeared roughly normal. There was a slight pattern of reading spring performance residuals plotted against the reading fall performance as a predictor; thus, a higher order (squared) term for the reading fall performance was examined for the ANCOVA model; however, the residual plot did not show a difference; thus, the higher order of the reading fall performance was not retained. The plots of standardized residuals against standardized predicted values were also checked for the six outcomes. These graphs show that the assumptions of linearity and homoscedasticity have been adequately met.

Table 4: ANCOVA tests of between-subjects effects for fall math performance.

Source	Type III Sum of Squares	F	Sig.	Partial Eta Squared
Corrected Model	331425.933[a]	257.359	.000	.234
Intercept	1975517.351	19942.350	.000	.645
Gender	173.172	1.748	.186	.000
Location Type	1389.231	3.506	.007	.001
Race	23129.845	77.830	<.001	.021
Region	7253.440	24.407	<.001	.007
School % FRPL	8618.732	87.004	<.001	.008
SES	117448.075	1185.609	<.001	.098
Error	1086604.607			
Total	12306769.367			
Corrected Total	1418030.539			

a. R Squared = .234 (Adjusted R Squared = .233).
NOTE: F = F-value as the Mean Square Regression divided by the Mean Square Residual; Sig. = p-value associated with the F-value; SES = Socioeconomic Status; FRPL = Free or Reduced-Price Lunch.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11(ECLS-K), Fall 2010 and Spring 2011.

Math Fall Performance The result of ANCOVA with students' math fall performance as the outcome is shown in Table 4 and Table 5. The ANCOVA results agreed with the regression tree results that students' SES is the strongest predictor in terms of the effect size partial eta squared and showed that students' racial group also played an important role. The linear regression results show that students from schools with different % free or reduced-price lunch had distinct math performance. However, the regression tree results showed that school % free or reduced-price lunch was only a strong indicator for the 10.15% for students in the upper SES scale range (0.260, 0.590]. Though regions remained a statistically significant indicator in the linear regression result, compared with other factors, it did not explain enough variance to show up in the best fitting two-level regression tree results.

Math Spring Performance The result of ANCOVA with students' math spring performance as the outcome is shown in Table 6 and Table 7 in the supplementary material. The ANCOVA results agreed with the regression tree results that students' math fall performance is the strongest predictor in terms of the effect size partial eta squared and showed that students' racial group and SES also played an important role. The linear regression results show that students from different regions had distinct math performance. However, the regression tree results showed that regions were only a strong indicator for the 10.00% for students in the upper math fall performance scale range (41.034, 46.679] and a relatively strong indicator (i.e., the third subdivision) for the 10.00% for students in the math fall performance scale range (36.458, 41.034] and for 8.62% White, Hispanic, and Asian students in the lower math fall performance scale range (33.489, 36.458]. Though school % free or reduced-price lunch remained a statistically significant indicator in the linear regression result, compared with other factors, it did not explain enough variance to show up in the best fitting three-level regression tree results. In contrast, though gender was not statistically significant in the ANCOVA result, the regression tree results showed

Table 5: ANCOVA parameter estimates for fall math performance.

	B	Std. Error	t	Sig.
Intercept	34.958	.439	79.559	.000
Female	−.251	.190	−1.322	.186
City, Midsize & Small	−.360	.350	−1.027	.304
Suburb	−.878	.329	−2.672	.008
Town	−1.004	.440	−2.283	.022
Rural	−1.194	.357	−3.346	<.001
Black	−2.996	.318	−9.431	<.001
Hispanic	−2.728	.269	−10.146	<.001
Asian	2.674	.389	6.879	<.001
Midwest	1.117	.327	3.412	<.001
South	1.747	.306	5.704	<.001
West	−.319	.322	−.989	.323
School % FRPL	−.039	.004	−9.328	<.001
SES	4.924	.143	34.433	<.001

NOTE: B = Unstandardized Coefficients; Std. Error = Standard Errors associated with the coefficients; t = t-value used in testing the null hypothesis that the coefficient is 0; Sig. = 2 tailed p-value used in testing the null hypothesis that the coefficient is 0; SES = Socioeconomic Status; FRPL = Free or Reduced-Price Lunch.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Kindergarten Class of 2010-11(ECLS-K), Fall 2010 and Spring 2011.

that gender was a relatively strong indicator (i.e., the third subdivision) for 8.77% students in the top math fall performance scale range >46.679 as well as for 6.13% White and Asian students in the middle math fall performance scale range (30.619, 33.489].

Math Gain (Spring 2011 Performance – Fall 2010 Performance) The result of ANCOVA with students' math gain as the outcome is shown in Table 8 and Table 9 in the supplementary material. The ANCOVA result showed that White students' math gain was statistically significantly different from Black and Asian students', but not from Hispanics'. The regression tree results were in line with the linear regression result by grouping White students and Hispanic students together and then Black and Asian students together in terms of math gain. Though school % free or reduced-price lunch and two out of three region variables (i.e., Midwest and West) were statistically significant according to the ANCOVA result, the regression tree result showed that school % free or reduced-price lunch and the region factors had different subgroup associations with students' math gain. The region factor was only a strong indicator for White and Hispanic students. The school % free or reduced-priced lunch indicator was the strongest indicator for Black and Asian students.

Reading Fall Performance The result of ANCOVA model with students' reading fall performance as the outcome is shown in Table 10 and Table 11 in the supplementary material. The ANCOVA results agreed with the regression tree results that students' SES is the strongest predictor in terms of the effect size partial eta squared and showed that students' racial group and region also played an important role. The ANCOVA results showed that female students had an average of 1.229-point advantage over male students. However, the regression tree re-

sults showed that gender was only a strong indicator for the 29.65% for students in the middle SES scale range (−0.430, 0.260] and a relatively strong indicator (i.e., the third subdivision) for 2.98% White and Asian students in the lower SES scale range (−1.090, −0.830] and 7.70% students from Midwest, Northeast, and South in the SES scale range (−0.830, −0.640]. Though school % free or reduced-price lunch remained as a statistically significant indicator in the linear regression result, compared with other indicators and factors, they only created the third subdivisions in subgroups. One of four locale type (i.e., Rural) was statistically significant in the linear regression model but did not appear in the regression tree result.

Reading Spring Performance The result of ANCOVA with students' reading spring performance as the outcome is shown in Table 12 and Table 13 in the supplementary material. The ANCOVA results agreed with the regression tree results that students' reading fall performance is the strongest predictor in terms of effect size partial eta squared and showed that students' SES and racial group also played an important role. Location types were a strong predictor in the ANCOVA results, but did not show up in the final best fitting two-level regression tree results. Similarly, school % free or reduced-price lunch was statistically significant in the ANCOVA but did not appear in the regression tree result. The ANCOVA results showed that female students had an average of 1.003-point advantage over male students. However, the regression tree results showed that gender was only a strong indicator for the 10.00% students in the upper reading fall performance scale range (50.224, 54.173] and for the 10.00% students in the bottom reading fall performance scale range $\leqslant$33.963.

Reading Gain (Spring 2011 Performance – Fall 2010 Performance) The result of ANCOVA with students' reading gain as the outcome is shown in Table 14 and Table 15 in the supplementary material. The ANCOVA result showed that locale type factors were statistically significantly related to students' reading gain. The regression tree results were in line with the ANCOVA result showing that the locale type factor is the strongest indicator. Though school % free or reduced-price lunch and two out of three racial group factors (i.e., Black and Hispanic) were statistically significant according to the ANCOVA result, the regression tree result showed that school % free or reduced-price lunch and the racial group factors have different subgroup associations with students' reading gain. The racial group factor was only a strong indicator for students from suburbs and cities. The school % free or reduced-price lunch indicator was the strongest indicator for students from rural and town; it also created the third subdivision for Black and Hispanic students in large cities. Though gender was statistically significant, the regression tree result showed that gender was only a strong indicator for Black, Hispanic, and Asian students in suburbs and midsize and small cities as well as students in schools with less than 8.09% free or reduced-price lunch in rural and town – a total of 30.73% population.

All three region factors (i.e., Midwest, South, and West) were not statistically significant according to the ANCOVA result, with the South factor being "almost significant" ($p = 0.056$) as the region with the lowest students' reading gain. However, the regression tree result showed that the region factor created the third subdivision for rural and town schools with more than 34.44% free or reduced-price lunch, for White students in suburbs and midsize and small cities, and for White and Asian students in large cities, meaning that the region factor was a relatively strong predictor for those subgroups. Also, for White students in suburbs and midsize and small cities, students from South made a gain of 15.290, which was distinct from an average gain of 14.197 of West, Midwest, and Northeast combined. The ANCOVA result showed that the

adjusted R^2 was small .016, thus, the predictors together did not explain much variance in the outcome variable; other predictors in the dataset may be considered to be included for future analyses.

4.2 Summary of Comparison between Linear and Nonlinear Results of ECLS-K Data

When the mean absolute errors of the regression tree models were compared with the mean absolute errors of the ANCOVA models, they were similar in the case of fall reading performance and fall math performance; however, ANCOVA had a slightly smaller mean absolute error than the regression trees for spring reading performance and spring math performance; in contrast, regression trees had slightly smaller mean absolute error than ANCOVA for reading gain and math gain. That is because the fall performance was a strong predictor of spring performance; in the regression tree models, CHAID algorithm separated this dominant continuous predictor into ten bins, while ANCOVA fitted the linear regression between the continuous predictor and the continuous outcome; thus, ANCOVA had slightly better prediction accuracy. But when there was not a dominant continuous predictor linearly related to the continuous outcome, as in the case of reading gain and math gain, regression trees are likely to have better prediction accuracy. This matches James et al.'s (2017) comparison of linear models and trees. On the other hand, linear models commonly suffer from multicollinearity, thus making it impossible to interpret the coefficients separately (Field, 2013). In contrast, the regression tree results made the interaction between predictors much easier to interpret.

The results have shown that the patterns illuminated by the regression trees differ across the subject areas (i.e., reading and math) and between the performance levels and achievement gains. The math and reading performances of the upper, middle, lower end of kindergarteners on the SES scale were variously related to different factors (i.e., race, region, and gender). For reading gain related, important factors included some of the same factors but also depend on location types (i.e., cities, suburbs, towns, and rural areas). The picture differed for math gain because racial groups performed differently based on their schools' characteristics – % FRPL and geographic designation. Rather than attempting to identify the significance levels of predictors uniformly affecting the whole population, the case study elucidates the complexity of predictors' roles in subgroup populations. Regression trees have the advantage of finding the best predictors for subgroups, by choosing the partition that is optimal for the particular step of the tree growing instead of being optimal for the entire tree, and making it easier to interpret. It is important to know how subgroups perform and make gains in different subject areas before pedagogies and interventions can be tailored to address students' needs.

Because achievement data are often strongly linearly related to continuous predictors – such as prior achievement or students' SES like as in this case study – linear regression models may have the advantage of prediction accuracy over regression trees. Even so, we may first fit a linear regression model and use regression trees with other predictors on the residuals to illustrate the complex relationship between the predictors and the outcome variable. In addition, if prediction accuracy is the goal, ensemble methods such as random forests and bagging can be used. Single trees have a clear disadvantage over ensemble methods in terms of prediction accuracy with a prediction accuracy of single trees, on average, being 10% less than ensemble methods (Loh, 2014). When interactions are not complex and can be correctly identified by a single tree, a single tree and bagging can perform equally well. However, in cases where there are many predictors with complex interactions, random forests tend to outperform bagging, while

both random forests and bagging tend to perform better than single trees (Strobl et al., 2009). However, single trees have a clear advantage in terms their interpretability (Loh, 2014).

In summary, when achievement data such as achievement gain in our case study are not strongly linearly related to any continuous predictors, regression trees may make more accurate predictions than linear regression models and also produce results that are easier to interpret. Thus, this case study has shown that regression tree analyses can be utilized with observational or survey data in education to display complex patterns between predictors and outcomes using a data-driven approach that would be obscured by trying to model the entire population as a whole. Similar analyses can be done with other predictors to enrich our understanding of the social context of early childhood education. Parallel studies can also be conducted with the 1998 ECLS-K database to identify changes in early childhood education.

5 Discussion

5.1 A More Complete Picture Provided by Using Tree-Based Methods

Quantitative research in education often follows a deductive tradition – researchers go into the field with a focused investigation starting with a research question and hypothesis and use the data either to fail to reject the null hypothesis or to reject the null hypothesis leading to evidence to support the alternative hypothesis. While the hypothesis testing is useful in reaching a decision or a conclusion about a research problem, it has its limitations as O'Dwyer and Bernauer (2013) state "following the scientific method by stating a hypothesis prior to conducting a study can limit researchers' ability to notice evidence that emerges to support competing hypotheses about how attributes or variables are associated with each other. That is, focusing on testing preconceived hypotheses can lead the researcher to miss important opportunities for future theory generation" (p. 65). CRT is a machine learning technique similar to the inductive process used by qualitative research and allows a pattern recognition from data. This provides an opportunity to complement the quantitative social scientific hypothesis testing tradition by giving researchers a complete picture and enabling them to generate new theories.

If the goal is to achieve prediction accuracy, tree-based methods can also serve as a benchmark predictor, so that the prediction accuracy achieved by simpler parametric models can be used to compare with the prediction accuracy of the tree-based methods. If the simpler parametric models can achieve the same level of prediction accuracy and can be easily interpreted, then the simpler parametric models should be used. However, if not, and variables of high importance in the tree-based methods are not present in the simpler parametric models, it will shed light on the fact that "relevant nonlinear or interaction effects may be missing in the simpler model which may not be suited to grasp the complexity of the underlying process" (Strobl et al., 2009, p. 19). In addition, tree-based methods can be used as a mechanism for variable selection – a smaller number of relevant predictors can be chosen from the full list of variables and are used subsequently in parametric models in a two-sage approach on a new set of data (Strobl et al., 2009).

5.2 Potentials and Limitations for Tree-Based Methods Application

In a linear model with the very common case of multicollinearity, it is impossible to interpret the coefficients separately and one is limited to simply making predicted values for the entire set of predictors. The regression tree can be viewed one step at a time. The difference in interpretability

is amplified when the linear model has interaction terms. In this case, CRT can be used to explore data and produce easily interpreted models. When similar cases are grouped into each node in CRT, it can describe data by showing the number and percentage of cases in each node, the mean, and the standard error, and the range or section of the variables to partition the outcome variable. However, when there is a strong linear relationship between one of the predictors and the response, in other words, one of the predictors is continuous and a dominant one, and the outcome variable is continuous, the tree may fit slightly worse than the linear model because the continuous predictor is broken into discrete bins with CHAID algorithm.

Many educational achievement data sets use sampling weights on each observation. Though developments have been made to use CRT while accounting for the weights or sampling design, it requires familiarity with certain statistical software and programs. Some large-scale assessment databases such as National Assessment of Educational Progress use plausible values as students' performance outcomes. Software such as "AM" or "EdSurvey" can conduct linear regression models and comparisons using plausible values. However, statistical programs that allow researchers to use plausible values for CRT analysis are not readily available.

5.3 Future Directions of Tree-Based Methods Application in Educational Research

Ma (2018) has proposed several hybrid statistical modeling approaches that can be used to exercise the full capacity of CRT. First, CRT can be used with HLM models to study longitudinal data or conduct multilevel CRT analyses. When analyzing longitudinal data, HLM can depict the rate of change over some time, while CRT can categorize subjects into different rates of change based on subject characteristics. When conducting multilevel analyses, by integrating CRT and HLM, the influence of the implicit data structure, which are the associations between the outcome variable and predictors depending on subject characteristics, can be revealed at the subject and cluster levels. Second, CRT can also be used for meta-analysis in ways that effect sizes can be categorized based on empirical studies' characteristics so that the interplays of study features can be unveiled. Third, CRT can also be used when there is more than one dependent variable. Categorical dependent variables can produce more refined categories based on categories of each categorical variable; the new categorical variable as a result of this procedure can be used as an outcome for CRT analyses. Two dependent variables with at least one being continuous can be entered into CRT analyses with the continuous dependent variable as the first independent variable to partition the root node and the other variable as a dependent variable (Ma, 2018). In the case of achievement data, because they are often strongly linearly related to continuous predictors such as prior achievement or students' socioeconomic status, we may consider a hybrid approach of linear models and regression trees by first fitting a linear regression model with the dominant continuous predictors and then using regression trees with other predictors on the residuals to illustrate the complex relationship between the predictors and the outcome variable.

Note on Data Restrictions

The data used in this paper is marked as restricted use by the National Center for Education Statistics (NCES). NCES required that all percentages in the tables provided be rounded to two decimal places for data security, and sample sizes to be rounded to the nearest ten. Further,

synthetic data are provided in the supplementary material. Access to the restricted use ECLS-K:2011 data can be applied for using the procedures at: https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2013060.

Supplementary Material

The supplementary material includes the following files: (1) README: a brief explanation of all the files in the supplementary material; (2) synthetic data files; (3) code files; (4) supplemental files for the manuscript – a. supplemental tree file: an expanded overview of CRT method, and b. supplemental tables and figures file: additional ANCOVA result tables and regression tree figures for the outcome variables.

References

Baker B (2001). Can flexible non-linear modeling tell us anything new about educational productivity? *Economics of Education Review*, 20(1): 81–92.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984). *Classification and Regression Trees*. CRC press.

Cheadle J (2008). Educational investment, family context, and children's math and reading growth from kindergarten through the third grade. *Sociology of Education*, 81(1): 1–31.

Cooper C, Crosnoe R, Suizzo M, Pituch K (2010). Poverty, race, and parental involvement during the transition to elementary school. *Journal of Family Issues*, 31(7): 859–883.

Field A (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage.

IBM Corp (2021a). *IBM SPSS Modeler, Version 18.3*. IBM Corp., Armonk, NY.

IBM Corp (2021b). *IBM SPSS Statistics for Windows, Version 28.0*. IBM Corp., Armonk, NY.

James G, Witten D, Hastie T, Tibshirani R (2017). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Jeon M, De Boeck P (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, 48(3): 1070–1085.

Jeon M, De Boeck P, van der Linden W (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics*, 42(4): 467–490.

Kass GV (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2): 119–127.

Kuhn M, Johnson K (2013). *Applied Predictive Modeling*, volume 26. Springer.

Ledolter J (2013). *Data Mining and Business Analytics with R*. John Wiley & Sons.

Loh W-Y (2014). Fifty years of classification and regression trees. *International Statistical Review*, 82(3): 329–348.

Ma X (2005). Growth in mathematics achievement during middle and high school: Analysis with classification and regression trees. *Journal of Educational Research*, 99(2): 78–86.

Ma X (2018). *Using Classification and Regression Trees: A Practical Primer*. Information Age Publishing, Inc.

Mulligan GM, Hastedt S, McCarroll JC (2012). *First-Time Kindergartners in 2010–11: First Findings from the Kindergarten Rounds of the Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K: 2011) (NCES 2012-049)*. U.S. Department of Education. National Center for Education Statistics, Washington, DC.

O'Dwyer LM, Bernauer JA (2013). *Quantitative Research for the Qualitative Researcher*. SAGE publications.

Rupp AA, Garcia P, Jamieson J (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, 1(3–4): 185–216.

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.

Tourangeau K, Nord C, Lê T, Sorongon AG, Hagedorn MC, Daly P, et al. (2015). *Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-K: 2011). User's Manual for the ECLS-K: 2011 Kindergarten Data File and Electronic Codebook, Public Version (NCES 2015-074)*. U.S. Department of Education. National Center for Education Statistics, Washington, DC.

Yan X, Su X (2009). *Linear Regression Analysis: Theory and Computing*. World Scientific.