A Data Mining Approach to Value-Added Models

Arturo Valdivia* Sharon L. Lohr†

Abstract

Many policymakers propose linking administrative decisions related to teachers to their effectiveness. Since teacher effectiveness cannot be measured directly, many researchers use value-added models to apportion changes in student achievement to the teachers and schools who have taught them. Several models, based on a linear mixed model framework, have been proposed and used for assessing value added by teachers: students' test scores are used as outcome variables and teachers' contribution is treated as random effect. The value-added score for a teacher is the empirical best linear unbiased predictor in the linear mixed model. However, the linear mixed model formulation has certain limitations, among them, its rigid structure. To address this, we use random forest. We introduce new variable importance measures and also use the existent measures to rank teacher effects. In addition, comparisons of traditional linear mixed model and random forest results are presented. We show that the random forest results may be more accurate when the linear model is misspecified. It is possible to use this approach as a complementary tool to linear models.

Key Words: Value-added models, random forest, variable importance measure, teacher effects

1. Introduction

Much of the debate about educational reform has centered on teacher and school accountability. Attempts to measure the teacher's influence on student achievement has been an interest in the scientific community for several decades (Hanushek, 1971; Bryk and Weisberg, 1976; Hanushek, 1979). Programs have been implemented in specific states or school districs to account for school and/or teacher effects since the early 1990s, such us The Tennessee Value-Added Assessment System (Sanders, Saxton, and Horn, 1997). However, since the most recent reauthorization of The Elementary And Secondary Education Act, The No Child Left Behind Act of 2001, a major emphasis has been placed on setting standards that each teacher must meet in order to be considered highly qualified. As a consequence, many states and school districts have adopted or are in the process of adopting models intended to measure teacher effects on student achievement.

A number of models have been proposed and are currently used for assessing value added by teachers and schools. Most of these models, henceforth called traditional VAMs, are either special cases of a general mixed model described in McCaffrey, Lockwood, Koretz, and Hamilton (2004) or extensions of it (McCaffrey and Lockwood, 2010; Mariano, McCaffrey, and Lockwood, 2010). In these models, students' test scores are used as outcome variables, while the contributions of teachers are treated as random effects. Hence, the value-added score for a teacher is obtained as the predicted value of the random effect.

The appropriateness of the use of VAMs in education is an ongoing debate (Stewart, 2006; Rothstein, 2009, 2010; Briggs and Domingue, 2011; Kinsler, 2012). This study approaches one potential limitation of traditional VAMs: the rigid structure of the model. Specifically, this limitation arises because the linear model structure only includes covariates that are explicitly included in the model; typically, few interactions are considered and nonlinearity is typically only considered through quadratic terms. It might be possible that a certain teacher is more effective with a certain group of students, but that situation can

*School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287-1804
†Westat, 1600 Research Blvd., Rockville, MD, 20850

only be taken into account in the VAM if the corresponding interaction is modelled in advance. If this is not the case, the VAM would be misspecified. To address this limitation we work with data mining methods, in particular random forest (Breiman, 2001), a supervised ensemble method of bagging or bootstrap aggregation. The advantage of using this method is that, as opposed to the traditional VAMs, no structure is predetermined. Therefore, in principle, any possible effect would be taken into account when obtaining results.

This work compares the information about teacher effect obtained using traditional VAM methodology and data mining techniques. For the former, we use two models: the covariate adjustment model and the gain score model. For the latter, we work with random forest. In the linear mixed models, the teacher effects are obtained using the empirical best linear unbiased predictors (EBLUPs). In random forest, there is no methodology that produces teacher effects. Rather, variable importance measures are used to obtain a ranking of teacher effects. Additionally, new importance measures are developed based on the random forest internal structure. Comparison of the alternative methods are obtained based on simulations that assume scenarios for both correctly specified and misspecified linear mixed models.

Section 2 presents the background literature for the linear mixed models and data mining methods used in this study. In Section 3, we describe existent variable importance measures, introduce new measures, and describe how comparisons between linear mixed model results and variable importance measures are made. The simulation study is describe in Section 4. Selected results are presented in Section 5. Concluding remarks are presented in Section 6.

2. Background

We provide a brief description of the VAMs and data mining techniques used in the following sections. In this study, VAMs are centered on teacher contributions and assume teacher effects are random effects while effects associated to other covariates (e.g. gender, rural or urban housing, free and reduced lunch status) are considered fixed. The data mining techniques used in this study are based on random forest (Breiman, 2001) and rankings of teacher effects are obtained using variable importance measures.

2.1 Value-added models

2.1.1 Covariate Adjustment Model (CAM)

The CAM assumes a single cohort of students in two contiguous years or grades, $t = 1,2$, where $t = 1$ corresponds to the first grade of the study. We assume there are N students, K_2 teachers in grade 2, and K_1 teachers in grade 1. The dependent variable is students' scores in year 2 of the study. The model uses scores in year 1 as one of the covariates in the model. In addition, teacher (random) effects in the model are considered. Formally, and letting the superscript c denote the parameters from the CAM, we have:

$$y_{i2} = \delta^c y_{i1} + \boldsymbol{\beta}^{c\prime}\mathbf{x}_i + \mathbf{b}^{c\prime}\mathbf{z}_i + \varepsilon^c_{i2} \tag{1}$$

where $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{iP})$ is the vector of covariates for student i, $\boldsymbol{\beta}^c = (\beta^c_0, \ldots, \beta^c_P)$ is the vector of fixed effects with intercept β^c_0, δ^c is the slope relating the year-2 score to the year-1 score, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK_2})$ is the vector specifying which teacher effects are associated with the year-2 score of student i, and $\mathbf{b}^c = (b^c_1, \ldots, b^c_{K_2})$ is the vector of teacher random effects. We assume that $\mathbf{b}^c \sim \mathbf{N}(\mathbf{0}, (\sigma^c_\tau)^2\mathbf{I})$. Similarly $\boldsymbol{\varepsilon}^c_2 = (\varepsilon^c_{12}, \ldots, \varepsilon^c_{N2}) \sim \mathbf{N}(\mathbf{0}, (\sigma^c)^2\mathbf{I})$. Moreover, $\boldsymbol{\varepsilon}^c_2$ is independent of $\mathbf{b}^c$. The predicted values of $\mathbf{b}^c$ will be the value-added

scores for the year-2 teachers. Note that teacher-level covariates can be included in this model by using the indicator variables in $\mathbf{z}_i$.

2.1.2 Gain Score Model (GSM)

The construction of GSM is similar to the one for CAM. The main difference is that the dependent variable is now the difference of year 2 and year 1. Specifically, we have $y_i^g = y_{i2} - y_{i1}$ and:

$$y_i^g = \boldsymbol{\beta}^{g\prime}\mathbf{x}_i + \mathbf{b}^{g\prime}\mathbf{z}_i + \varepsilon_i^g \tag{2}$$

where $\mathbf{x}_i$ is student i's vector of covariates related to fixed effects. $\boldsymbol{\beta}^g = (\beta_0^g, \ldots, \beta_P^g)$ is the vector of fixed effects with β_0^g the overall mean. $\mathbf{b}^g = (b_1^g, \ldots, b_{K_2}^g)$ is the vector of teacher random effects. The assumptions are similar to those in CAM; $\mathbf{b}^g \sim \mathbf{N}(\mathbf{0}, (\sigma_\tau^g)^2\mathbf{I})$ and $\boldsymbol{\varepsilon}^g = (\varepsilon_1^g, \ldots, \varepsilon_N^g) \sim \mathbf{N}(\mathbf{0}, (\sigma^g)^2\mathbf{I})$, $\boldsymbol{\varepsilon}^g$ is independent of $\mathbf{b}^g$.

2.1.3 Teacher Effects

The teacher contributions to student achievement are estimated using the empirical best linear unbiased predictor (EBLUP):

$$\hat{\mathbf{b}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{3}$$

where $\mathbf{V} = \mathbf{cov}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$ is the covariance matrix of $\mathbf{y}$. For the CAM (1) we have: $\mathbf{y}' = (y_{12}, ..., y_{N2})$, $\mathbf{X} = (\mathbf{x}_1^*, ..., \mathbf{x}_N^*)'$ with $\mathbf{x}_i^* = (y_{i1}, \mathbf{x}_i)$, $\boldsymbol{\beta} = (\delta^c, \boldsymbol{\beta}^c)$, $\mathbf{b} = \mathbf{b}^c$, $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_N)'$, $\mathbf{G} = ((\sigma_\tau^c)^2\mathbf{I})$, and $\mathbf{R} = ((\sigma^c)^2\mathbf{I})$. Similarly, for the GSM (2), we have: $\mathbf{y}' = (y_1^g, ..., y_N^g)$, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)'$, $\boldsymbol{\beta} = \boldsymbol{\beta}^g$, $\mathbf{b} = \mathbf{b}^g$, $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_N)'$, $\mathbf{G} = ((\sigma_\tau^g)^2\mathbf{I})$, and $\mathbf{R} = ((\sigma^g)^2\mathbf{I})$.

2.2 Random Forest

First, we briefly describe random forest. For a more detail description see Breiman, Friedman, Olshen, and Stone (1984); Breiman (2001); Loh (2008); Hastie, Tibshirani, and Friedman (2009).

Some comments about the notation used in what follows: N and $i = 1, ..., N$ represent the observations in our data set. We loosely speak of N for regression trees and random forest, since the methods often require a split of the original data set in two subsets that are known as training (or learning) data set and test data set. This is common practice in data mining, since the training data set is used to build the model for prediction and the test data set is used to find certain statistics or measures, such as levels of accuracy, impurity, overfitting, etc. This division could be dynamic as is the case in cross-validation. When needed, we use $\mathscr{L}_N = (Y_i, X_{i1}, ..., X_{iP}); i = 1, ..., N_1$ and $\mathscr{T}_N = (Y_i, X_{i1}, ..., X_{iP}); i = 1, ..., N_2$ for the learning data set and the test data set, respectively, where $N_1 + N_2 = N$.

Similarly in random forest, the set of data used on each tree is obtained using bootstrapping, a sample with replacement from the original data set that also has N observations, some of which are sampled more than once while others are not sampled at all. For ease of notation and when no confusion arises we will use N and $i = 1, ..., N$ for both the original data set and each set obtained through bootstrapping. For every tree obtained in this way, the observations that are not included form the *out-of-bag* samples (OOB). OOB samples are meaningful in the development of several VIMs methods discussed later. When needed, we use $\mathscr{B}$ and $\mathscr{B}^C$ for the bootstrap and the OOB samples, respectively.

Regression trees and forests are built using covariates. Let X_p for $p = 1, \ldots, P$ denote each one of the covariates in the model, and let X_{ip} denote the value of that covariate for

student i, $i = 1,\ldots,N$ so that $\mathbf{X}_{i\cdot} = (X_{i1},\ldots,X_{iP})$. In general, covariates may be any feature associated with student i; demographics such as gender or race/ethnicity, socio-economic status indicators such us free or reduced lunch program, urban or rural housing, or, most importantly for this study, the teachers associated with the student. Our study of VIMs centers on the importance of the individual covariates describing which teachers taught the students, so we set aside the first M of the covariates to be the teacher indicators. For $m = 1,\ldots,M$, let $X_{im} = 1$ if teacher m instructed student i and 0 otherwise. These categorical variables thus represent the presence or absence of a particular teacher for each student.

In the following, T and sometimes R, represent a specific tree. In addition, for tree T, D_T and J_T represent the number of internal and terminal nodes, respectively, or simply D and J if no confusion arises. Similarly, d_{q_1}, $q_1 = 1,\ldots,D$ and j_{q_2}, $q_2 = 1,\ldots,J$ represent individual internal and terminal nodes, respectively, or simply d and j when no confusion arises. When the relationship between the children and their parent node is needed, we use d_ℓ and d_r for the left and right children nodes of d. We use indicator variables in a variety of contexts, that assign 1 when the condition is met and 0 otherwise. Thus, $\zeta_{kd} = 1$ indicates that observation k landed in node d, $\phi_{pd} = 1$ represents when covariate p is used as the splitting variable in node d, $\omega_{jk} = 1$ reflects when observation k corresponds to terminal node j, $\omega_{dk} = 1$ reflects when observation k corresponds to internal node d, and $\nu_{mk} = 1$ reflects when student k has been taught by teacher m.

2.2.1 Random Forest and Variable Importance Measures

Random forest (Breiman, 2001)is the average prediction obtained from a collection of regression trees, themselves built through the random generation of a subset of attributes in the data.

Random forest has become a popular method in several research fields. Its appeal lays in its predictive accuracy that is comparable to the best machine learning methods. In particular, random forest performs well when the structure of the underlying model is nonlinear, the number of covariates is very large, covariates are highly correlated, and/or complex interactions are present among covariates. Additionally, random forest is used as a method for variable selection via the use of variable importance measures (VIMs).

3. Obtaining Variable Importance Measures

Variable importance measures or importance scores are measures used to determine the relative contribution that each covariate has in predicting the dependent variable. VIMs have been used for variable selection and variable ranking in several fields during the last decade. But, as far as we know, this is the first time they have been used to assess relative contributions of teachers. We describe the existent methods used to obtain VIMs, comment about potential advantages and limitations, and explain how they give an indication of the relative teacher effectiveness. Furthermore, we propose new methods to obtain VIMs and discuss their relevance in the context of VAMs.

3.1 Decision Trees Variable Importance Measure

Breiman et al. (1984) propose an importance measure for decision trees based on the estimated improvement in squared error loss that a variable has in the internal nodes of the tree. To understand this we need to present a few definitions. The squared error loss at node d is defined as $e_d^2 = \sum_{Y_i \in \mathscr{T}_N} (Y_i - \bar{Y}_d)^2 \zeta_{id}$, where $\bar{Y}_d = \frac{1}{\sum_{Y_i \in \mathscr{L}_N} \zeta_{id}} \sum_{Y_i \in \mathscr{L}_N} Y_i \zeta_{id}$. That is, the sum of the squared differences between the values of the outcome variables in the test data set that

arrive to node d and the mean of the outcome variables from the training data set in node d. The improvement in squared error loss is defined as $\iota_d^2 = e_d^2 - (e_{d_\ell}^2 + e_{d_r}^2)$. Formally, if T is a decision tree with $D-1$ internal nodes (the root of the tree is not considered), the importance measure for variable X_p in tree T is defined as:

$$VI_p^2(T) = \sum_{d=1}^{D-1} \iota_d^2 \phi_{pd} \tag{4}$$

where

$$\phi_{pd} = \begin{cases} 1 & \text{if variable } X_p \text{ is used as splitting variable in internal node } d \\ 0 & \text{otherwise.} \end{cases}$$

That is, the squared relative importance of covariate X_p is the sum of improvements in squared error loss for every node where X_p is used as the splitting variable. Intuitively, there are two components that influence on this variable importance measure. First, the covariates that are found as splitting criterion closer to the root of the tree are potentially more important than those covariates closer to the leaves of the tree. This happens because more observations are considered in nodes closer to the root and therefore the improvement in square error loss tends to be greater. Second, for a particular node, how different the means of the children nodes are determines how importance the covariate is relative to that node.

Notice that a covariate that has a large number of categories, has a larger number of possible splitting points. If this number is larger, relative to other covariates, this covariate will tend to have a larger variable importance values. Therefore, this measure may be biased toward covariates with larger number of categories.

For a random forest with R trees, we use the average of these measures obtained for every single tree: $VI_p^2 = \frac{1}{R}\sum_{r=1}^{R} \sum VI_p^2(T)$. Although random forest will correct some bias given the random selection of covariates for every node, variables with larger number of categories, and therefore larger number of splitting points, would still be favored towards selection, and will obtain a larger variable importance measure.

3.2 Permutation Accuracy Importance.

Permutation accuracy importance (PAI), introduced by Breiman (2001), is obtained for each covariate p as the difference in prediction accuracy between the original OOB data set and its permuted version, where the permutation occurs only for covariate p. Formally, if $f(\cdot, \mathscr{B}, \Theta_T)$ is the random forest solution for tree T when the bootstrapping sample is $\mathscr{B}$, the variables considered for splitting at each node δ are given by Θ_T, and the OOB sample is $\mathscr{B}^C$, the estimated prediction accuracy is:

$$\Lambda(\mathscr{B}, \Theta_T) = \frac{1}{|\mathscr{B}^C|} \sum_{i:(\mathbf{X}_{i\cdot}, Y_i) \in \mathscr{B}^C} (f(\mathbf{X}_{i\cdot}, \mathscr{B}, \Theta_T) - Y_i)^2 \tag{5}$$

where $|\mathscr{B}^C|$ is the number of observation in the OOB sample. The prediction accuracy for the random forest with R trees is given by the average prediction accuracy of all the trees. For convenience, let us express this result in terms of covariate p: $\Lambda(\mathbf{X}_{\cdot p}) = \frac{1}{R}\sum_{r=1}^{R} \Lambda(\mathscr{B}, \Theta_r)$. We then permute the values of covariate p in the OOB samples to create a new sample for each tree. That is, the data set values are the same for all the observations and covariates, except those corresponding to covariate p. Those are randomly reasigned in a different order. We called this new sample $\mathbf{X}^*_{\cdot p}$, and we obtain $\Lambda(\mathbf{X}^*_{\cdot p})$. Finally, the variable importance measure based on permutation accuracy importance is $PAI = \Lambda(\mathbf{X}^*_{\cdot p}) - \Lambda(\mathbf{X}_{\cdot p})$. The

intuition behind this method is that if a covariate is important, the permutation should produce a large gap between the prediction accuracy of the original OOB samples and the one obtained from permuting variable p.

Similar to the limitations of *decision trees VIM*, PAI is not reliable when covariates are of different type (e.g. continuous and categorical, qualitative and quantitative), the variables are on a different scale of measurement, or the variables have different number of categories. This happens because a covariate with a larger number of categories relative to other covariates, will tend to have a better prediction accuracy and a larger difference with the permuted version. Hence, it will be biased towards covariates with larger number of categories. Also, the PAI overestimates the importance of correlated covariates; variables that are not important might be considered much more relevant, because they might be highly correlated with other covariates. To address the problem of different type of covariates and different scales of measurement, Hothorn et al. (2006) present a method for growing trees based on a conditional inference framework. To address the problem of correlation among covariates Strobl, Boulesteix, Kneib, Augustin, and Zeileis (2008) introduce the conditional variable importance.

3.3 Shrinkage in Variable Importance Measures

When obtaining VIMs based on random forest, there is a shrinkage effect similar to the one described for the linear mixed model random effects. Intuitively, this effect is present because in random forest, variables are selected randomly for each node, and therefore even when a covariate specific random effect is much larger than others, this covariate could only potentially appear on the nodes where it has been selected. Similarly to EBLUPs for teacher effects, the VIMs obtained from random forest take into account the entire data, not only the teacher's own students. Notice, however, this shrinkage effect will be affected by the number of students each teacher has. Teachers with fewer students do not appear as frequently in the trees. This occurs from two sources. One source is that only a subsample of the observations is considered for each tree in random forest, and a teacher with fewer students might have even fewer students in certain trees or sometimes not students at all. Additionally, most VIMs are determined by the number of observation affecting each node in which the covariate is used as splitting variable, and the number of times that covariate appears in the tree. Therefore, teachers with fewer students might not only be considered less important than teachers with more sutdents, their estimates might also be less accurate than the estimates of teachers with more students. These two opposing effects, the shrinkage effects as well as how the number of students per teacher influence the VIMs, is addressed in Section 5.

3.4 A New Approach to Variable Importance Measures

As previously mentioned, several approaches to determine variable importance measures have been proposed and developed. These approaches have advantages and limitations. The main characteristic of most approaches is that they use the accuracy of predictions in normal and altered conditions in order to determine how important each covariate is. While some of these methods have shown empirical success, they are based in strong assumptions about the distribution of the covariate p, the independence between this covariate and the dependent variable, the differences in variable type among covariates, etc. Even the conditional approach cannot fully take into account the possible correlation between covariates. Furthermore, these methods seem not to take full advantage of the structure of the regression trees and random forest. Empirical results have shown that the accuracy of random

forest preditions is comparable to the best learning machine methods; however, variable importance measures have not had the same level of success.

With this motivation, we propose a new approach to VIMs. We present measures that are constructed based entirely on the structure of regression trees and random forest and are not dependent on prediction accuracy differences. This is the fundamental difference between the proposed new measures and those previously found in the literature. It could present advantages, since any assumptions and considerations should be made before growing the trees and no posterior assumptions are necessary, as is the case with PAI. At the same time, it could present a limitation, since the quality of the measures is a reflection of our understanding of the tree structure. We describe first the proximity matrix, which is used as the base for developing the new VIMs.

3.4.1 Proximity Matrix in Random Forest

A by-product of regression trees and random forest is the proximity matrix. This is an $N \times N$ symmetric matrix, where every cell represents the proportion of ocurrences that the observation represented by the row position belongs to the same terminal node as the observation represented by the column position. If we consider the proximity matrix for a single tree, it is a matrix of zeros and ones, where a coordinate with a *one* indicates that two observations, the first determined by the row position and the second by the column position, share the same terminal node.

3.4.2 Node and Teacher Proportions

Every existent VIMs requires the use of the outcome values of the dependent variable from two sources. The first source is used while building the tree, since the sum of square error differences of the outcome values are used to determine the covariates and covariates' values used for splitting each node. The second source comes from using outcome values to construct the different types of VIMs, for example via comparisons of predicted with observed values or improvements in square error loss. It seems that if we could avoid the use of the second source, and use instead the already existent tree structure to obtain VIMs, we could potentially avoid the limitations or biases exclusively introduced in our VIMs by the second source.

This is the main motivation to propose new VIMs. The *node* and *teacher* proportions are VIMs that take advantage exclusively of the structure of the tree and outcome values are not used in their construction. In addition these VIMs are specifically designed for covariates with characteristics present in the VAM context; in particular, covariates representing teachers. These covariates are binary variables that represent the presence (1) or absence (0) of that covariate in the corresponding observation (student). Both new measures try to capture from different perspectives how the final configuration of each tree in the random forest gives information about the importance of each covariate. There are different approaches that could be used to obtain this information. We have decided to depart from the information of the tree obtained from proximity matrices for each tree in the random forest, and pinpoint unique characteristics for each covariate.

Specifically, we continue assuming N students and M teachers. The relative importance of covariate (teacher) m on observation (student) k in tree r is given by:

$$V_{mrk}^{node} = v_{mk} \sum_{j=1}^{J_r} \omega_{jk} \frac{\sum_{i=1}^{N} \omega_{ji} v_{mi}}{\sum_{i=1}^{N} \omega_{ji}} \tag{6}$$

where

$$\nu_{mk} = \begin{cases} 1 & \text{if covariate (teacher) } m \text{ is present in observation (student) } k \\ 0 & \text{otherwise} \end{cases}$$

and

$$\omega_{jk} = \begin{cases} 1 & \text{if observation (student) } k \in \text{terminal node } j \\ 0 & \text{otherwise} \end{cases}$$

Equation 6 represents the relative contribution of teacher m in student k in tree r. It is obtained as the ratio of the sum of weights of teacher m's students in students k's terminal node over the sum of weights of all students in student k's terminal node. Furthermore, we find the total contribution of teacher m in tree r, as the average of all the individual contributions V_{mrk}^{node} for $\{k : 1,\ldots,N \text{ and } \nu_{mk} = 1\}$ in tree r, and the VIM for teacher m is the average of all those contributions on every tree. Formally, we have:

$$VIM_m^{np} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{\sum_{i=1}^{N}\nu_{mi}}\sum_{k=1}^{N}V_{mrk}^{(s)} \tag{7}$$

We call (7) the *node proportion VIM* for student m. In what follows, we denote by VIM_{np} the measures obtained following this procedure.

An alternative approach to measure the relative importance of covariate m is given by:

$$V_{mrk}^{teach} = \nu_{mk}\frac{\sum_{j=1}^{J_r}\omega_{jk}\sum_{i=1}^{N}\omega_{ji}\nu_{mi}}{\sum_{j=1}^{J_r}\omega_{jk}\sum_{i=1}^{N}\nu_{mi}} \tag{8}$$

Equation 8 measures also the relative contribution of covariate m on observation k in tree r, but using a slightly different approach. It is the ratio of the sum of weights of observation k's terminal node where covariate m is present, over the sum of all observations where covariate m is present in tree r. Similarly, the VIM for covariate m is the average of V_{mrk}^{teach} for all observations k in tree r where m is present, and all trees r, $r : 1,\ldots,R$. We have:

$$VIM_m^{tp} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{\sum_{i=1}^{N}\nu_{mi}}\sum_{k=1}^{N}V_{mrk}^{(s)} \tag{9}$$

We call (9) the *teacher proportion VIM* for student m, and denote by VIM_{tp} the measures obtained following this procedure.

3.5 Comparing $VIMs$ with $VIM_{\ell m}$

The VIMs and measures of variable importance based on linear mixed model estimates (henceforth $VIM_{\ell m}$), are not directly comparable. The VIMs determine the relative importance of each teacher effect. That is, the VIMs produce a ranking of teacher effects while the $VIM_{\ell m}$ produce teacher effect estimates. However, we can also obtain a ranking of importance of teacher effects from the $VIM_{\ell m}$, by taking the absolute value of the teacher effects estimates. Additionally, in our simulations we know by design the exact influence that each covariate has on the dependent variable. Therefore, using the absolute value of the true teacher effects, we can make comparisons between VIMs and $VIM_{\ell m}$.

One of the limitation of only producing a ranking of teacher effects is that we do not produce the direction of the effects. From a practical point of view, this is not optimal since one of the motivations of using VAMs is to determine if a teacher effect is positive or negative.

Although there are ways to overcome this limitation, our main motivation is to determine the adequacy of these measures first. And we do so, without estimating the direction

of the effects for two reasons. First, as mentioned previously, the proposed VIMs, *node* and *teacher proportion*, do not use outcome variables in their construction, but only the structure of the tree. If a direction for the measure was introduced based on outcome variables, the new measures proposed would have an additional influence that could potentially mask the differences between the existent methods to obtain VIMs and the proposed ones. Second, using our knowledge of the true effects in the simulations, we could assign the correct sign to the VIMs. However, the results obtained following the latter procedure lead us to the same conclusions as the absolute value approach.

We therefore make comparisons between the ranking of teacher effects using Spearman's rank correlation coefficients for our comparisons.

4. Simulation Study

In this section, we introduce the study design and the description of the factors used.

4.1 Data structure and design

Several factors are manipulated in this study, including:

*1. The number of teachers.*10, 20, 40, or 100 teachers.

2. The number of students per teacher in each group (SpT_1/SpT_2). We divide the data in two groups and assign to each teacher within a group the same number of students. Teacher in different groups could have different number of students, where SpT_ℓ is the number of students per teacher in group ℓ. The ratios of students per teacher considered are: $\frac{12}{12}$, $\frac{24}{24}$, $\frac{36}{36}$, $\frac{36}{12}$, or $\frac{30}{18}$.

3. Ratio of teacher variance /student variance (σ^2_τ/σ^2). Choices used in manipulation of this factor include 1, 2, 5, and 20. For example, σ^2_τ/σ^2 = 5 would indicate that the teacher variance is five times as large as the student variance.

4. The number of trees in random forest. 100, 500, 1000, 2000, and 3000 trees are considered.

5. Model type. Two types of models considered: the *covariate adjustment (CAM)* and the *gain score model (GSM)* (see Section 2).

6. Model specifications. For each of the model types (*CAM* and *GSM*), a family of four models is generated. The first is the baseline model described in section 2. For the other model specifications the following extensions of the baseline models are given.

The extension of the *covariate adustment model* is given by:

$$y_{i2} = \delta^c y_{i1} + \boldsymbol{\beta}^{c\prime}\mathbf{x}_i + \mathbf{b}^{c\prime}\mathbf{z}_i + \sum_{j=1}^{P}\sum_{k=1}^{K_2}\lambda^c_{jk}x_{ij}z_{ik} + \sum_{j\neq\ell}\sum_{k=1}^{K_2}\lambda^c_{j\ell k}x_{ij}x_{i\ell}z_{ik} + \varepsilon^c_{i2} \tag{10}$$

The extension of the *gain score model* is given by:

$$y^g_i = \boldsymbol{\beta}^{g\prime}\mathbf{x}_i + \mathbf{b}^{g\prime}\mathbf{z}_i + \sum_{j=1}^{P}\sum_{k=1}^{K_2}\lambda^g_{jk}x_{ij}z_{ik} + \sum_{j\neq\ell}\sum_{k=1}^{K_2}\lambda^g_{j\ell k}x_{ij}x_{i\ell}z_{ik} + \varepsilon^g_i \tag{11}$$

The baseline model represents the family of simulations that assume no interaction effects. For *CAM* model this holds true when λ^c_{jk} and $\lambda^c_{j\ell k}$ in (10) are equal to zero for all $j,\ell : 1,...,P$ and $k : 1,...,K_2$ and all the assumptions hold. In this case the linear mixed model is correctly specified, and the random effects estimates are the EBLUPS. For the simulations we consider four covariates associated with fixed effects: the *prescore* is obtained from a normal distribution with mean 75 and variance 21, *gender* is a binary variable obtained from a binomial distribution with probability of success $p = 1/2$, *urban* or *rural*

housing, another binary variable obtained from a binomial variable with a probability of having urban housing of $p = 0.4$, and *free and reduced lunch program* also a binary variable with probability $p = 0.8$ of being part of the program. The associated fixed effects used in the simulations for these variables are .5, .2, 3, and -5, respectively. Furthermore, an overall mean of 50 is considered. The *error* variance is set equal to 1, and therefore the ratio $\sigma_\tau^2/\sigma^2 = \sigma_\tau^2$, the teacher variance.

The *good teacher - bad teacher* model represents the family of simulations that does not account for interactions effects and keeps most of the assumptions of the baseline model except the one related to the distribution of $\mathbf{b}$. ($\mathbf{b}^c$ for CAM and $\mathbf{b}^g$ for GSM). Specifically, these models are constructed with only two teachers having large effects in the model, one positive and the other negative, while the rest of the teachers have no effects. This model specification is used, because it does not meet the assumptions of the linear model and might be better suited for data mining techniques. For the simulations, the positive effect is set at $1.5 * \sigma_\tau^2$ and the negative effect at $-1 * \sigma_\tau^2$. Here, σ_τ^2 is used only for setting the good and bad teacher and does not represent the teacher variance. We use the same set of covariates associated with fixed effects and values used in the respective baseline model.

The simple interaction model represents the family of models that include simple interactions effects between one covariate associated with a fixed effect, x_{ij}, and another covariate associated with a random effect, z_{ik}. In *CAM*, this is represented in (10) by having at least one $\lambda_{jk}^c \neq 0$ for $j : 1, ..., P$ and $k : 1, ..., K_2$. An interaction effect of 10 is considered in the simulations, for half of the teachers randomly determined when a student, taught by one of those selected teachers, lives in the rural area. In *GSM*, simple interactions are modeled when at least one $\lambda_{jk}^g \neq 0$ for $j : 1, ..., P$ and $k : 1, ..., K_2$ in (11).

The complex interaction model is the family of models that include interactions among three covariates, two of them associated with fixed effects and one associated with random effects. For *CAM*, this is represented in model (10) by having at least one $\lambda_{j\ell k}^c \neq 0$ for $j, \ell : 1, ..., P$, $j \neq \ell$, and $k : 1, ..., K_2$. In the simulations, we randomly determined half of the teachers to be susceptible to this interaction effect, and the interaction studied corresponds to students living in an urban area, belonging to the free and reduced lunch program, and being taught by one of these teachers. We considered and interaction effect of 20. In *GSM*, the complex interaction is indicated when at least one $\lambda_{j\ell k}^g \neq 0$ for $j, \ell : 1, ..., P$, $j \neq \ell$, and $k : 1, ..., K_2$ in (11) (Note that for the simulations, we used the same values as in *CAM*).

4.2 Procedures and analysis

The simulations are based on a factorial design yielding a total of 3200 combinations. Five VIMs are considered:
a) the absolute value of the linear mixed model random effect estimates, $VIM_{\ell m}$,
b) the VIM based on the PAI, denoted by VIM_1,
c) the VIM based on the improvement in square error loss, denoted by VIM_2,
d) the *node proportion*, VIM_{np} and
e) the *teacher proportion*, VIM_{tp}.

For each experiment, 100 replicates are obtained and the Spearman's rank correlation is computed between the absolute value of the *true teacher effects* and the VIMs, for each replicate. The results for each experiment are expressed as the correlation mean taken over all the replicates. In the next Section, selected results are reported.

5. Results

For limitations of space, we only present selected results for CAM (additional results can be obtained by request). Four different scenarios are considered, each representing a row of graphs within a figure. Scenarios under consideration include: CAM_1 as the baseline model, CAM_2 as the *good teacher-bad teacher* model, CAM_3 as a simple interaction model, and CAM_4 as a complex interaction model (see Section 4 for a complete description).

Figure 1 plots the correlations means for the five measures across the number of teachers (x-axis) when the number of students per teacher is either 12 (left column charts) or 24 (right column charts) and the ratio of teacher variance over student variance is 2.

When the number of students per teacher was 12 in CAM_1, the $VIM_{\ell m}$ yielded higher correlations than the remaining measures. As the number of students per teacher increased, the data mining VIMs tended to improve in performance, although $VIM_{\ell m}$ still outperformed the rest. It is to note that the correlations stabilized at 40 teachers, suggesting little change in CAM_1 due to an increase in the number of teachers from 40 to 100 teachers. All the VIMs based on random forest had similar performance when the number of teachers was 40 or higher, however VIM_1 and VIM_2 slightly outperformed the *teacher* and *node* proportions, when the number of teachers was 10 or 20. When the number of students per teacher was 24, all the VIMs based on random forest improved and reduced considerably the gap towards the $VIM_{\ell m}$ results in comparison to the case with 12 students per teacher. Most of the other conclusions remained the same as in the case with 12 students.

When the number of students per teacher was 12, in CAM_2, the $VIM_{\ell m}$ only produced slightly better results when the number of teachers was 10 or 20; when the number of teachers was 40 or higher, the performance was practically the same for all five measures. Moreover, measures tended to suffer in performance relative to the results obtained in CAM_1, in particular as the number of teachers increased. These results seem reasonable in view of the shrinkage estimate existent in both the linear model and the VIMs estimates. The true teacher effect for most teachers is zero and only two teacher effects are away from zero. The larger the number of teachers, the stronger the shrinkage effect on all teachers, and the more difficult it is to recognize which teachers are those with nonzero effects. In consequence, the performance of all the VIMs suffered. When the number of students per teacher was 24, the $VIM_{\ell m}$ had the same performance as all the VIMs based on random forest for every studied number of teachers. The conclusions about the measures' performance for different number of teachers remained the same as with the 12 students per teacher case.

When the number of students per teacher was 12 in CAM_3, similar patterns to that of CAM_1 were found. The $VIM_{\ell m}$ produced better results than all the other measures; all four measures based on random forest performed similarly when the number of teacher was 40 or greater. When the number of teachers was 10 or 20, VIM_1 and VIM_2 outperformed the *teacher* and *node* proportions. Additionally, when the number of students per teacher was 24, the results and trends in performance remained the same. As in CAM_1, the gap in performance between $VIM_{\ell m}$ and the VIMs based on random forest was smaller. Further, similar results were obtained when other simple interactions were considered. It seems the linear mixed model estimates adjust quite well to the presence of a unique, however large, effect.

Results with Complex Interactions. Scenarios where complex interactions are modeled, CAM_4, present important results for this study. First, we focus on the case where the number of students per teacher was 12 (this chart is presented in the lower left corner of Figure 1). We observed that $VIM_{\ell m}$ outperformed all the VIMs based on random forest only when the number of teachers was 10. When the number of teachers was 20, $VIM_{\ell m}$ outperformed VIM_1 and VIM_2, but obtained a similar performance to that of the *teacher* and *node* propor-

tions. When the number of teacher was 40, only VIM_2 did not outperform $VIM_{\ell m}$. Finally, when the number of teacher was 100, all the VIMs based on random forest outperformed $VIM_{\ell m}$. These results are important. The $VIM_{\ell m}$ was obtained based on a linear mixed model that does not account for the complex interactions, and the Spearman correlation between the absolute value of true teacher effects and the absolute value of EBLUPs is lower than the correlation obtained when the model was correctly specified (without interactions). Furthermore, all the VIMs based on random forest performed better, since random forest accounted for all the possible interactions. Most importantly, the two proposed VIMs, the *node proportion* and the *teacher proportion*, outperformed the rest of the measures when the number of teachers was 20 or greater.

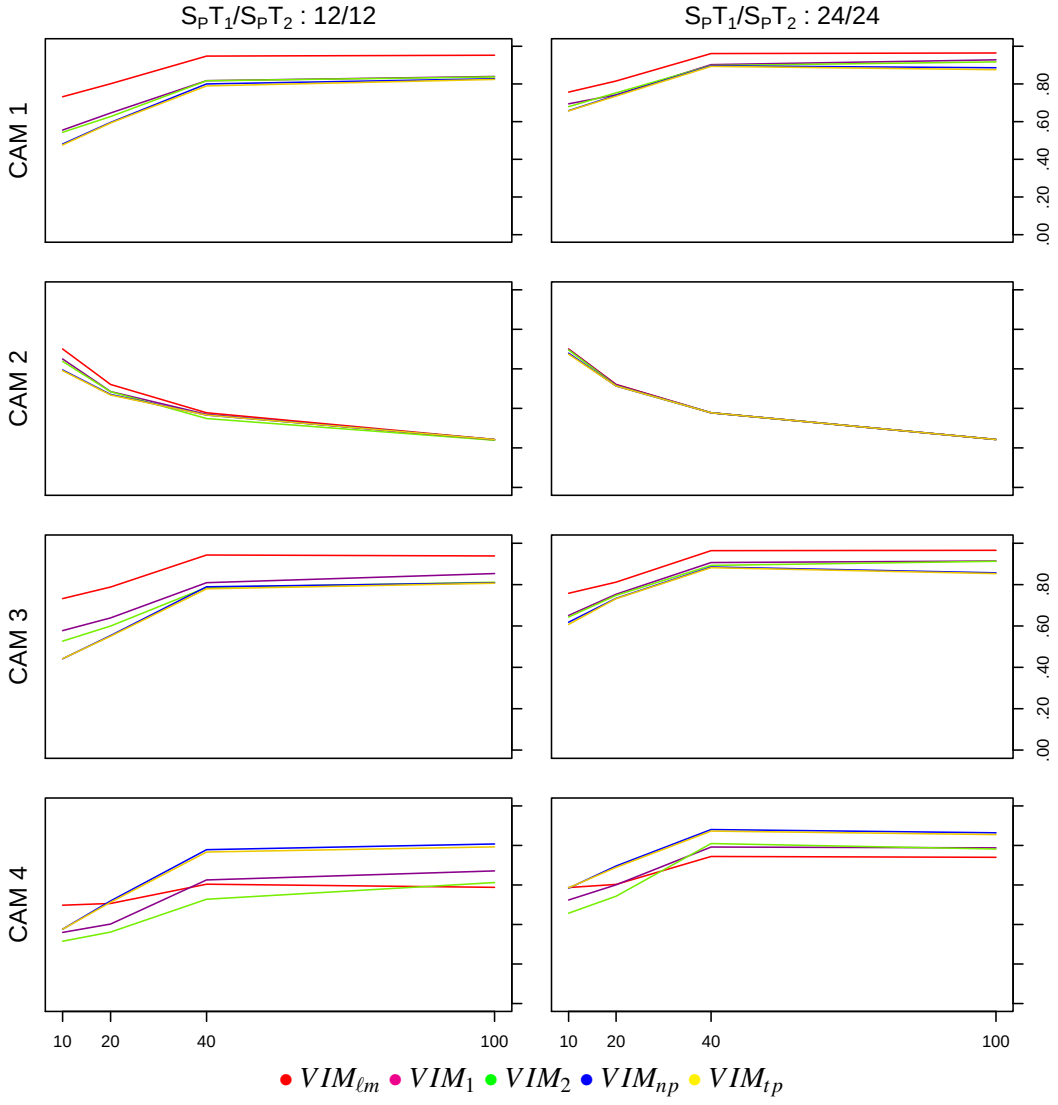

Figure 1: Mean correlation between the VIMs and the absolute value of true teacher effects when the number of teachers varies for different CAM models and different student per teacher ratios. $\sigma^2_\tau/\sigma^2 = 2$

Intuitively, *teacher* and *node proportions* capture better the complex structure of the model, or more precisely, the random forest captures the complex structure of the model and these measures reflect more accurately this information. For example, a component of the *node proportion* is obtained as the ratio of students in a terminal node that have

been taught by a specific teacher over the total number of students in that terminal node. The complex structure of the model is taken into account by random forest when generating those terminal nodes, and thus, this ratio includes that complex structure. As a result, the measure improves when the number of covariates (teachers) increases; random forest can obtain more information about the complex structure of the data, and thus, the *node proportion* can better reflect the teachers' importance. By contrast, when the number of teachers is small, there is less oportunity to fully exploit the tree structure and the performance of the *node proportion* is not as strong. On the other hand, VIM_1 and VIM_2 use additional information based on the outcome values, which seems to add certain noise to the measures' accuracy when complex interactions are present. When the number of students per teacher was 24, the results and conclusions were very similar. In this situation, *teacher* and *node proportions* perform at least as well as $VIM_{\ell m}$ when the number of teachers was 10, and better than any other measures when the number of teachers were 20 or greater.

In comparing the results from all four *CAM* families, the *node proportion* and *teacher proportion* produced very consistent estimates for CAM_1, CAM_3, and CAM_4. By contrast, $VIM_{\ell m}$ produced similar results for CAM_1 and CAM_3, but the performance was greatly affected by complex interactions, CAM_4. Furthermore, in the case of 24 students per teacher, the measures' performance improved slightly over that of 12 students per teacher. However, in most situations, the relative performance of $VIM_{\ell m}$ did not change with respect to VIMs based on random forest.

Similar results were obtained for *GSM*. While the patterns of measures' performance remained relatively similar, the main difference between *GSM* and *CAM* was in the magnitude of the results. The mean correlations in *GSM* were not as high as those found in *CAM*. It would seem that the covariate *prescore* played an important role in determining the ranking of teacher effects.

6. Concluding Remarks

This work studies value-added models from a new perspective. We use data mining methods, in particular random forest, to evaluate the accuracy of random effects estimates obtained from linear mixed model formulations, in situations where these formulations are correctly specified or misspecified. Although the focus here is on the value-added models in education, the proposed methods could extend to any area where value-added models are used.

When a linear mixed model is correctly specified, there is no better random effect estimate than the empirical best linear unbiased predictor. However, the ranking of random effects obtained using variable importance measures is not far behind, and in several scenarios, this ranking is almost as good as the one obtained with the linear model. On the other hand, when the linear mixed model is misspecified, the results obtained using variable importance measures based on random forest may produce a more accurate ranking of teacher effects. This happens in particular when the true model presents complex interactions that are not considered in the linear model specification.

The obtained results are important in many respects. First, a large difference between the ranking obtained via variable importance measures and via linear model estimates should signal that the linear mixed model is misspecified. Second, although this method could help determine model misspecification, the exact simple and complex effects are not known, and alternative methods need to be developed for this purpose. We are currently working on methods that help determine simple and complex interactions based on the random forest structure.

In terms of variable importance measures, we have proposed two new methods. These

methods rely exclusively on the trees structure, and do not use the outcome values or any other additional information, as it is the case in the traditional variable importance methods. In reported and unreported results, we have found that in situations with complex interactions, the new variable importance measures consistently outperformed the existent variable importance measures. Furthermore, we believe the new proposed measures could be improved. These measures use the trees structure of random forest only to the extent of the final configuration of the terminal nodes. However, the information of the internal nodes could also be considered, and it is an area where additional research is needed.

Limitations of this study are based on the extent of our simulations. We have obtained results based on the covariate adjustment model and the gain score model. Currently, we are working on the complete persistence and generalized persistence model, proposed by Mariano et al. (2010).

In conclusion, this study addresses an area unexplored until now in the theory and practice of value-added models and linear mixed models. Although additional work is needed to confirm and extend the results of this study, the initial findings are encouraging. Future work could address the issue of estimation of teacher effects with alternative data mining methods.

References

Leo Breiman. Random forest. *Machine Learning*, 45(1), 2001.

Leo Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

Derek Briggs and Ben Domingue. Due diligence and the evaluation of teachers: A review of the value- added analysis underlying the effectiveness rankings of los angeles unified school district teachers by the los angeles times. National Education Policy Center, 2011. URL "http://nepc.colorado.edu/publication/due-diligence.

Anthony S. Bryk and Herbert I. Weisberg. Value added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics*, 1:127–155, 1976.

Eric A. Hanushek. Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review*, 60:280–288, 1971.

Eric A. Hanushek. Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, 14:351–388, 1979.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

Josh Kinsler. Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics*, 3(2):333–362, 07 2012.

W.-Y. Loh. Classification and regression tree methods. In F. Ruggeri, R. Kenett, and F. W. Faltin, editors, *Encyclopedia of Statistics in Quality and Reliability*, pages 315–323. Wiley, Chichester, UK, 2008. URL http://www.stat.wisc.edu/ loh/treeprogs/guide/eqr.pdf.

Louis T. Mariano, Daniel F. McCaffrey, and J. R. Lockwood. A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, 35:253–279, 2010.

D. F. McCaffrey, J. R. Lockwood, D. M. Koretz, and L. S. Hamilton. Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 2004.

Daniel F. McCaffrey and J. R. Lockwood. Missing data in value-added modeling of teacher effects. *Annals of Applied Statistics*, page to appear, 2010.

Jesse Rothstein. Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, 4(4), 2009.

Jesse Rothstein. Teacher quality in education production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1):175–214, 2010.

William L. Sanders, Arnold M. Saxton, and Sandra P. Horn. *The Tennessee Value-Added Assessment System: A Quantitative, outcomes-based approach to Educational Assessment*, pages 137–162. Corwin Press, Thousand Oaks, CA, 1997.

Barbara Elizabeth Stewart. Value-added modeling: The challenge of measuring educational outcomes. *Carnegie Corporation of New York*, 2006.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9 (307), 2008.