The added value of more accurate predictions for school rankings

Fritz Schiltz, Paolo Sestito, Tommaso Agasisti, Kristof De Witte

Paper prepared for the 16th Conference of IAOS
OECD Headquarters, Paris, France, 19-21 September 2018

Wednesday, 19/09, 18h00: Poster Session

Fritz Schiltz
fritz.schiltz@kuleuven.be
KU Leuven

Paolo Sestito
paolo.sestito@bancaditalia.it
Banca d'Italia

Tommaso Agasisti
tommaso.agasisti@polimi.it
Politecnico di Milano

Kristof De Witte
kristof.dewitte@kuleuven.be
KU Leuven,
Maastricht University

The added value of more accurate predictions for school rankings

DRAFT VERSION 06/09/2018

Prepared for the 16th Conference of the
International Association of Official Statisticians (IAOS)
OECD Headquarters, Paris, France, 19-21 September 2018

ABSTRACT

School rankings based on value-added (VA) estimates are subject to prediction errors, since VA is defined as the difference between predicted and actual performance. We introduce the use of random forest (RF), rooted in the machine learning literature, as a more flexible approach to minimize prediction errors and to improve school rankings. Monte Carlo simulations demonstrate the advantages of this approach. Applying the proposed method to Italian middle school data indicates that school rankings are sensitive to prediction errors, even when extensive controls are added. RF estimates provide a low-cost way to increase the accuracy of predictions, resulting in more informative rankings, and more impact of policy decisions.

Keywords: value-added, school rankings, machine learning, monte carlo

1 Introduction

School rankings are increasingly being used as a means to strengthen accountability in the education sector (Nunes et al., 2015). Value-added (VA) models are considered a best practice to rank schools and have been adopted in, among others, the UK, Hong Kong, and the USA (Leckie and Goldstein, 2017). Apart from school rankings, VA models are being used to evaluate teachers (Backes et al., 2018), school principals (Branch et al., 2012), and even physicians (Fletcher et al., 2014). Estimating VA is a high-stakes statistical exercise, as rankings based on VA estimates often determine personnel decisions or school closure (Angrist et al., 2017).

Two caveats are worth noting with respect to school rankings based on VA estimates. First, earlier research has argued that nonrandom selection of students into classes and schools (sorting) biases VA estimates (Rothstein, 2009). Including controls can partially account for this bias, in those cases where sorting is on observables (Koedel et al., 2015). Let alone data issues, it is generally difficult to tell which, and how variables influence student sorting.[1] Second, VA estimation requires predictions, as they indicate the difference between actual and predicted performance. Hence, VA estimates are subject to prediction errors. Particularly, nonlinear interactions between important inputs in the education production function might result in unrealistic predictions when conventional linear estimates are used. Moreover, this issue of prediction errors remains in place, even when all relevant sorting variables are included.

This paper proposes the use of a 'random forest' as an alternative approach to estimate school VA and to obtain rankings. Random forests add flexibility by capturing nonlinearities and complex interactions (Breiman, 2001). A recent trend towards machine learning in economics advocates such models for predictions, as they may allow for more effective ways to model complex relationships (Mullainathan and Spiess, 2017; Varian, 2014). Especially when modelling the education production function, discontinuous relationships and nonlinear interaction effects are more naturally accommodated by a random forest. This machine learning approach does not require prior knowledge on the education production function (inside the 'black box'). Given the same set of variables, this added flexibility results in more accurate predictions.

We use Monte Carlo simulations to demonstrate that random forest estimates reflect more closely the VA of schools compared to conventional methods, resulting in more reliable school rankings. We then illustrate the benefits of the proposed approach using Italian middle school data. In addition to the availability of rich data, the Italian case is particularly interesting as there is an ongoing policy debate on the most appropriate statistics to publish as a guide for parents. This paper contributes to this public debate, which is also present in many other countries (for example, a new VA measure was recently introduced in the UK and has been heavily criticized (Leckie and Goldstein, 2017)).

Our simulations and empirical application indicate that school rankings based on conventional VA estimates are very sensitive to prediction errors emanating from restrictive functional form assumptions. More accurate VA estimates result in more informative rankings for parents, and more impact of policy decisions. Moreover, our results indicate that the improved accuracy from more flexible random forest estimates is comparable to the accuracy gains from adding more data in the conventional approach. This suggests a low-cost way to improve VA estimates, particularly when lim-

[1] Other studies have exploited data from lotteries to reduce bias in VA estimates (Deming, 2014; Angrist et al., 2017)

ited data is available. Similar gains in accuracy can be expected in other value-added contexts, for example when predictions are used at the teacher level rather than at the school level to evaluate teachers (Chetty et al., 2014; Hanushek and Rivkin, 2010). The proposed method is likely to be fruitful in other public sector activities as well, such as health and social services, where entities are ranked and evaluated based on value-added estimates.[2]

2 Empirical strategy

The value-added (VA) of a school measures how much better a school is doing than expected. Estimating school VA implies predicting individual student test scores and averaging the prediction errors for each school. A conventional approach is to predict test scores using a linear OLS regression of previous scores and students characteristics:[3]

$$A_i = \beta^{'}\mathbf{X}_i + v_i \tag{1}$$

$$\text{where } v_i = \mu_j + \varepsilon_i \tag{2}$$

with $A_{i,t}$ the test score (e.g. mathematics or reading) for student i, $\mathbf{X}_i$ the set of predictor variables, μ_j the effect of school j, and ε_i the unobserved error in scores, unrelated to the school VA. The VA of school j (μ_j), can then be obtained by averaging the prediction errors for school j:[4]

$$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N}(A_i - \hat{A}_i) \tag{3}$$

$$\text{where } \hat{A}_i = \beta^{'}\mathbf{X}_i \tag{4}$$

Schools that, on average, manage to help students achieve test scores A_i beyond their prediction $\hat{A}_i$ are considered to be adding value to the test scores, and vice versa. Student characteristics $\mathbf{X}_i$ and ε_i may be correlated with μ_j in the likely event that students self-select into schools. Accounting for such sorting behavior is the key challenge in obtaining unbiased VA estimates.

If all sorting variables could be observed, and are included in $\mathbf{X}_i$, the conventional OLS estimate of $\beta^{'}\mathbf{X}_i$ would potentially still lead to ε_i and μ_j being correlated due to anomalies in the education production function. For example, decreasing returns to scale not captured by (1), can result in erroneous predictions of $\hat{A}_i$, and hence $\hat{\mu}_j$. This source of bias in VA estimates can be reduced by increasing the accuracy of predictions $(A_i - \hat{A}_i)$.

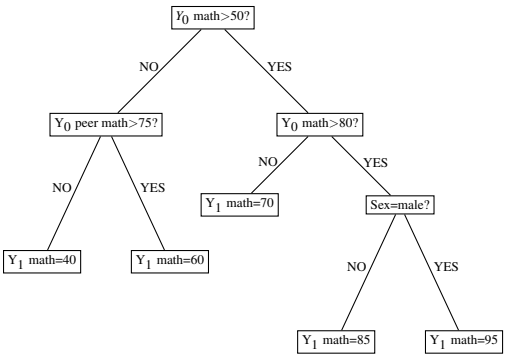
Using regression trees we can improve predictions for students, without the need to specify a functional form for the education production function. This can be done using the same set of variables, included in $\mathbf{X}_i$. For example, one could predict a student's test score with a regression tree using his or her previous test score(s) and the test scores of the student's classmates to account for peer effects. As in (1), the school fixed effect

[2] An R template code for other applications is available upon request.

[3] Our results are analogous for alternative specifications of the conventional model (linear-log, and higher degree polynomials), as the flexibility provided by random forests goes beyond these functional form adjustments. Note, however, that these adjustments are rather uncommon, with a linear specification of predictor variables being the standard functional form.

[4] Alternative approaches are often applied to mitigate bias in VA estimates of small schools, for example, by shrinking estimates towards the overall mean (Angrist et al., 2017; Guarino et al., 2015) or by using multilevel models. The latter approach has been adopted by the UK government to rank schools on VA (Leckie and Goldstein, 2017).

Figure 1: An example of a regression tree to predict mathematics scores in Y_1



Notes: Outcome variable is mathematics score in Y_1 [0-100]. The regression tree displays the fictitious result of recursive partitioning using Y_0 scores, peer scores in Y_0 and sex.

is not included when predicting $\hat{A}_i$ with regression trees. Instead, equivalent to the conventional approach, the school effect is obtained by averaging prediction errors for each school.

Regression trees can be seen as a set of rules resulting from recursive partitioning observations into groups, or 'leaves' - incorporating nonlinearities and interactions by construction (Breiman et al., 1984). These leaves are chosen to minimize the residual sum of squares (RSS) of the predictions. Figure 1 illustrates how regression trees add flexibility when predicting individual test scores. In this example tree, mathematics scores in Y_1 (here [0-100]) are predicted using individual and peers' scores in Y_0, and sex. In this fictitious example, a split at the mathematics score in Y_0 equal to 50 is best able to reduce the RSS compared to all other binary splits of all variables. Therefore, it is chosen as the first split, at the top of the tree. Along the same lines, an additional split on sex only reduces the RSS for high achieving students (Y_0 math > 80). As illustrated by this example, regression trees allow for nonlinear relationships between variables and the predicted outcome. Once the regression tree is built using recursive partitioning, the predicted outcome of a new observation then equals the average of the leaf where we end up by following the set of rules embodied by the tree. For example, a student with a math score in Y_0 above 50 and above 80 is predicted to score 85 if the student is a girl.

Using the nonlinear patterns identified by regression trees, predictions can be made for each student. Then, the added value of schools can be computed by comparing students' actual test scores, relative to the predicted test scores - i.e. the prediction errors. Averaging the prediction errors for each school yields the VA estimate, as in (3).

The accuracy of regression trees can be substantially improved by constructing them iteratively. A random forest (RF) constructs a large number of trees using randomly drawn samples and randomly drawn predictors as candidates at each split (Breiman, 2001; James et al., 2013). Corresponding parameters are ideally obtained using cross-validation to prevent overfitting. We will use random forests in the empirical application and Monte Carlo simulations to improve predictions, and hence, VA estimates and school rankings.

3 Data and specification

We construct school rankings for Italian middle schools by estimating each school's value-added for mathematics. We use data from the National Institute for the Evaluation of the Educational System of Education and Training (INVALSI). It contains extensive information on all Italian students and schools for the 2013 cohort. INVALSI resembles to the OECD PISA data, although data are collected for all students in Italy, and at different moments in time. It was designed in this way for the purpose of estimating the value-added of schools and to construct corresponding rankings. Every student is observed twice in the data: grade 5 data is collected at the end of primary school, and grade 8 data at the end of middle school. Hence, the change in mathematics test scores between grade 5 and grade 8 can be used to measure the added value of middle schools for mathematics.[5]

We first estimate a *baseline* model to predict grade 8 mathematics test scores including lagged test scores as the only predictor variables (i.e. in grade 5). This specification corresponds to VA in its most common form (Todd and Wolpin, 2003) - or the 'VA2' model used to compile school league tables in the UK. Using these predictions (and prediction errors), a VA estimate can be obtained for each school. We follow Lefgren and Sims (2012) by including lagged test scores for both mathematics and reading as predictor variables, to improve predictions of grade 8 test scores in mathematics. Next, we estimate a *final* model which can be seen as the 'contextual' VA used in the UK to obtain school rankings (Leckie and Goldstein, 2017), accounting for differences in student characteristics and peers, and hence reducing bias from student sorting. In particular, we include a set of student characteristics (immigrant status, sex, socio-economic status, grade repetition before grade 5), and peer characteristics, by averaging the same set of variables both at the class and school level. In addition, we also include the relative previous position of students in their class, and the relative previous position of students' classes in their schools. We do not claim to perfectly control for nonrandom selection of students into classes and schools, even though the INVALSI data allows a more complete set of controls than commonly included. To compare VA estimates and school rankings, we estimate both baseline and final models using a conventional OLS approach and using a random forest. When estimating the random forest, we set our parameters as follows: the number of trees (500), the number of observations per end node (15) and the number of variables as split candidates (1 and 24, for the baseline and final model). We chose this combination as it minimizes 10-fold cross-validation errors (a process often described as hyperparameter tuning). This way we maximize predictive power without overfitting, a common issue when using machine learning methods - see for example Mullainathan and Spiess (2017) for a more elaborate discussion.

4 Results and discussion

4.1 Monte Carlo simulations

In order to compare the ability of conventional and random forest estimates to reflect the school value-added (VA), we iteratively generate a sample of students and assign them to schools. First, we compare the accuracy of conventional and random forest predictions at the student level ($A_i - \hat{A}_i$). Second, we obtain VA estimates by averaging

[5] A more comprehensive description can be found in earlier studies using this dataset, see for example De Simone (2013, p.14) or Bertoni et al. (2013, p.66-67).

Table 1: Comparing accuracy of predictions

	Absolute error			Mean squared error	
Model:	RF	Conventional	Diff	RF	Conventional
baseline	20.52	23.78	-3.26***	716	979
final	9.48	22.54	-13.06***	151	869

Notes: Predicted variable is mathematics score in grade 8 (mean=198, SD=38). Predictor variables in baseline model are lagged test scores (grade 5) for mathematics and reading. Final model adds immigrant status, sex, socio-economic status, grade repetition before grade 5, class and school level averages of this set of variables, the relative previous position of students in their class, and the relative previous position of students' classes in their schools. *Diff* indicates the difference in absolute prediction errors. *** indicates significance at 1%.

these prediction errors. We then compare the VA estimates for each school relative to the true value added in our simulation ($\mu_j - \hat{\mu}_j$). Finally, we compare the school rankings obtained from the conventional and the RF approach to the true ranking. As detailed in the appendix of this paper, three parameters define the data-generating process: the effect size of school VA, the nonlinearity in the education production function, and the degree to which high ability students are sorted into high VA schools. Simulations over this set of parameters indicate that RF provides more accurate predictions of student test scores. Moreover, when the effect size of school VA is relatively modest, and the education production function is not strictly linear, we provide evidence that RF estimates are also better able to reflect the VA of schools, and are hence more informative about school rankings. This information gain is especially pronounced when students sort into schools.

We also explore more realistic settings were students are not only sorted on ability but also on demographics (i.e. high SES more likely assigned to better schools); or when VA effects of schools are heterogeneous across students (i.e. low SES students are affected more); or in the absence of peer effects, or when the conventional VA model accounts for nonlinearities in the EPF using higher degree polynomials. Our findings are robust to all these alternative scenarios, strengthening the case for the use of RF model in real life applications.[6]

4.2 Ranking Italian schools

Table 1 compares the accuracy of predicting individual mathematics test scores in grade 8 using INVALSI data, in terms of absolute errors and mean squared error (MSE).[7] Clearly, the random forest predictions outperform conventional predictions for both baseline and final models. Adding more data, i.e. going from baseline to final, reduces prediction errors. However, adding flexibility, i.e. going from conventional to RF, seems to reduce these errors even further, and significantly. The higher accuracy of RF predictions reveals the limited ability of conventional estimates to adequately capture the complex education production function, and casts serious doubt on accountability prescriptions based on such measures.

Next, we construct VA estimates from our predictions and rank Italian schools accordingly. Building on the findings from the Monte Carlo simulations, we set the

[6]Simulation results for these scenarios are summarily presented in Table A3. Alternative parameter combinations are available upon request.

[7]Our results are analogous for alternative definitions of school VA: (1) school median VA, (2) reading VA, (3) average of reading and mathematics VA.

Table 2: Identifying bottom and top schools

		Q25	
Model:	RF	Conventional	Diff
baseline	84.68	81.78	2.90***
final	100	85.75	14.25***
		Q75	
Model:	RF	Conventional	Diff
baseline	78.60	72.28	6.32***
final	100	81.72	18.28***

Notes: Percentages indicate the share of Italian schools classified as bottom (Q25) or top (Q75) by both benchmark rankings and the evaluated model using INVALSI data. See Table 1 for a description of predicted and predictor variables in baseline and final models. Benchmark rankings are those obtained from the final RF model. Bootstrapped SEs: *** indicates significance at 1%.

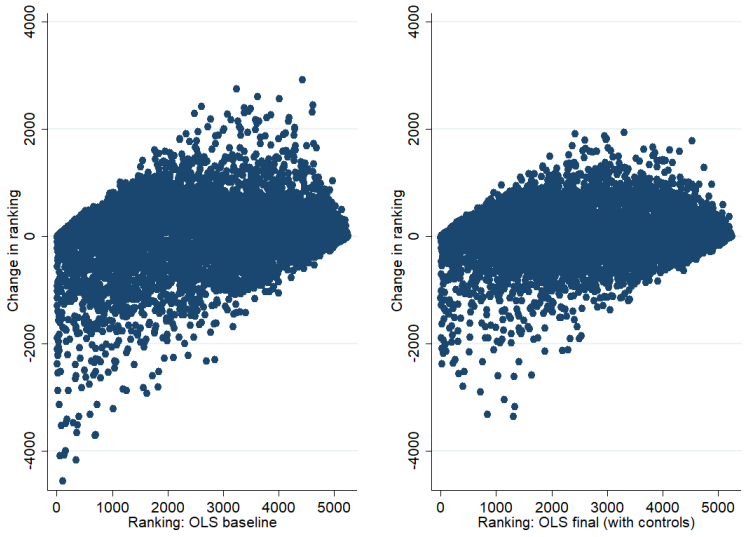
school ranking obtained from the final RF model as the benchmark ranking. In Figure 2, we compare rankings obtained from conventional estimates (baseline and final) to this benchmark. Clearly, major changes occur when rankings are based on RF estimates instead of conventional estimates of VA. These changes are especially pronounced in schools ranked at the bottom by conventional estimates, and even more so at the top. Also, including the extensive set of controls in the final specification appears to only partially resolve the diverging rankings. In the right hand side panel, after adding all controls to account for selection on observables, schools ranked highly by the conventional VA still experience large (downward) rank changes when compared to the RF rank.

Table 2 presents the share of schools correctly identified as ranked in the bottom or top quartile. We define 'correct' as a match with the classification obtained from the final RF, considering its ability to minimize prediction errors. For example, 82% of the schools ranked in the top quartile by the final RF estimate are also classified in this group using the final conventional estimate. As can be seen from Table 2, RF estimates of VA are significantly better at identifying low- and top-performing schools. This suggests a major advantage of ranking schools based on RF estimates when limited data is available.

For policy makers, school rankings can be particularly useful to identify best practices or to target low-performing schools. In practice, VA measures are used to rank schools and close down schools that end up at the bottom of this ranking. The impact of any such policy depends on the ability of rankings to identify schools at the bottom and at the top of the unknown VA distribution. Back-of-the-envelope calculations of closing the average school in the bottom quartile and enrolling its students in the average school in the top quartile indicate that achievement gains could be as large as 0.16 standard deviations (SD).[8] This is the policy impact when rankings obtained from the final RF estimates are used. However, when baseline conventional estimates of VA are used to obtain rankings, this effect reduces to 0.14 SD, as schools are being closed that are not actually in the bottom and students are sent to schools that are not actually in the top. A school closure policy based on RF estimates and limited data would yield the same benefits (0.15 SD) as a policy based on conventional estimates using the full set of controls. This implies that the policy impact can be increased by almost

[8]Following Angrist et al. (2017), we ignore possible transition effects such as disruption due to school closure, peer effects from changes in school composition, and other factors that might inhibit replication of successful schools.

Figure 2: Rank changes when improving predictions.



Notes: Schools ranked in terms of VA estimates: baseline conventional OLS estimates (left) and final conventional OLS estimates (right). We ranked all 5,249 Italian schools using their VA estimate based on INVALSI data. We obtain VA estimates for conventional and RF predictions by averaging the difference between actual and predicted scores, as in (3). The school ranked 1st exhibits the largest VA. Vertical axes indicate the change in rank when final RF estimates are used to obtain the ranking of schools. See Table 1 for a description of predicted and predictor variables in baseline and final models.

0.01 SD when extensive controls are added to the specification, and the impact can be increased by another 0.01 SD when flexibility is added to estimate VA and to rank schools. Although this effect appears negligible, it suggests that RF predictions provide an effective, low-cost way to improve rankings, irrespective of the data available.

5 Conclusion

For parents and policy makers, the main concern regarding school rankings is whether they provide a valid tool to compare school quality. However, since 'value-added' (VA) is defined as the difference between predicted and actual performance, prediction errors can result in biased rankings. This paper introduced random forests to estimate school VA, as this approach more naturally accommodates discontinuous relationships and nonlinear interaction effects in the education production function. Using Monte Carlo simulations we demonstrated that random forest estimates not only provide better individual predictions, but also provide a better approximation of school VA compared to conventional estimates, in nearly all parameter configurations. Starting from these findings, we compared rankings of Italian middle schools for random forest and conventional estimates. Clearly, rankings were strongly divergent, to an extent that could not be accounted for by including a set of controls at the individual, class, and school level. Finally, we provided back-of-the-envelope calculations to assess the impact of a hypothetical school closure policy in Italy, strictly based on rankings. Our calculations

indicate that the impact of this policy is increased by 0.01 standard deviations when random forest estimates are used to rank schools instead of conventional estimates.

When it comes to predictions, machine learning methods are prevailing in economics. They are often referred to as 'black box' methods due to their lack of transparency (Varian, 2014). This poses possible threats when implementing such models to guide education policies (e.g. school closure). However, similar claims can be made about conventional VA models - the UK government scrapped contextual value-added models in 2010 motivated by their lack of transparency, claiming it was 'difficult for the public to understand' (DfE, 2010, p.68). At a very basic level, regression trees are intuitive to explain a predictive process, as human decisions tend to follow a tree structured approach as well. There is a clear scope for further research in terms of communicating machine learning models more intuitively, e.g. in a visual manner. Nonetheless, considering the high stakes nature of school rankings, and VA estimates in general, improving accuracy may outweigh reduced transparency, as only small improvements can imply major changes for school principals and parents.

Acknowledgments

We are most grateful to Patrizia Falzetti and Paola Giancomo for providing access to the INVALSI data. We are also grateful to the editor, two anonymous referees, and seminar participants in Rome, Leuven, London and Lucca for many useful comments and remarks. This work was supported by the European Union's Horizon 2020 research and innovation programme [grant number 691676].

References

Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *Quarterly Journal of Economics*, 132(2):871–919.

Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, L. C., and Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, 62:48–65.

Bertoni, M., Brunello, G., and Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics*, 104:65–77.

Branch, G. F., Hanushek, E. A., and Rivkin, S. G. (2012). Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals. Technical Report w17803.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and regression trees*. CRC press.

Chetty, R., Friedman, J. E., and Rockoff, J. N. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9):2633–2679.

De Simone, G. (2013). Render unto primary the things which are primary's: Inherited and fresh learning divides in Italian lower secondary education. *Economics of Education Review*, 35:12–23.

Deming, D. J. (2014). Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review: Papers & Proceedings*, 104(5):406–411.

DfE (2010). The importance of teaching: The Schools White Paper.

Fletcher, J. M., Horwitz, L. I., and Bradley, E. (2014). Estimating the Value Added of Attending Physicians on Patient Outcomes. *NBER Working Paper Series*, w20534.

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P. N., and Wooldridge, J. M. (2015). An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures. *Journal of Educational and Behavioral Statistics*, 40(2):190–222.

Hanushek, E. A. and Rivkin, S. G. (2010). Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review*, 100(2):267–271.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge, Abingdon.

James, G., Witten, D., Hastie, R., and Tibshirani, R. (2013). *An introduction to statistical learning*. Springer, New York, 6 edition.

Koedel, C., Mihaly, K., and Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47:180–195.

Leckie, G. and Goldstein, H. (2017). The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'progress 8'. *British Educational Research Journal*, 43(2):193–212.

Lefgren, L. and Sims, D. (2012). Using Subject Test Scores Efficiently to Predict Teacher Value-Added. *Educational Evaluation and Policy Analysis*, 34(1):109–121.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nunes, L. C., Reis, A. B. B., and Seabra, C. (2015). The publication of school rankings: A step toward increased accountability? *Economics of Education Review*, 49:15–23.

Rothstein, J. (2009). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, 4(4):537–571.

Sacerdote, B. (2014). Experimental and Quasi-Experimental Analysis of Peer Effects: Two Steps Forward? *Annual Review of Economics*, 6(1):253–272.

Todd, P. E. and Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal*, 113(485):2–33.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *The Journal of Economic Perspectives*, 28(2):3–28.

Appendix: Monte Carlo simulations

In order to compare the ability of conventional and random forest estimates to reflect the school value-added (VA), we iteratively ($B = 100$) generate 10,000 student observations, grouped in 100 schools. First, we compare the accuracy of predictions at the student level, measured as the mean squared error (MSE). Second, we compare the MSE of VA estimates for each school relative to the true value added in our simulation. Finally, we compare the rankings obtained from the conventional approach and the RF rankings to the true ranking using rank order correlation coefficients (Spearman's ρ and Kendall's τ).

1 Data Generating Process

We define the data generating process (DGP) of student achievement in Y_1 as a function of previous test scores, peer test scores in Y_0, and the added value of schools. For each student, we calculate test scores in Y_1 as follows:

$$\begin{aligned} &\hat{M}_{1i} = f(M_0, M_0^p, \mu_s) + \varepsilon_i \\ &\text{with } \varepsilon_i \sim N(0,1) \end{aligned} \tag{1}$$

1.1 Functional form: α

The functional form underlying the DGP and connecting mathematics test scores in Y_0 and Y_1 is specified as a linear combination of a linear and nonlinear function. In particular:

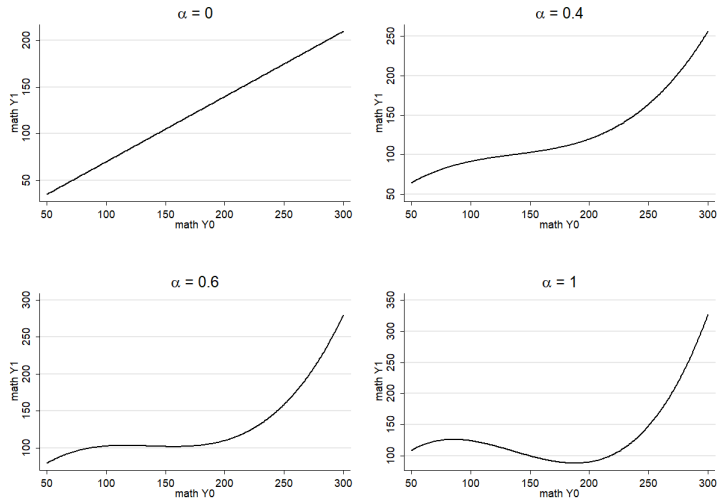
$$\begin{aligned} \hat{M}_{1i} =& (1-\alpha)[\beta_M(M_{0i} + \beta_p M_{0i}^p)] \\ &+ \alpha[\sum_1^k \beta_{Mk}(M_{0i} + \beta_p M_{0i}^p)^k] \\ &+ \mu_s + \varepsilon_i \end{aligned} \tag{2}$$

We set $k = 4$, obtaining a fourth degree polynomial function in the second part of (2). As this part is clearly a nonlinear function of M, α indicates the degree of nonlinearity in the education production function (EPF). For $\alpha = 0$, (2) reduces to a strictly linear specification, whereas $\alpha = 1$ imposes the polynomial functional form on the EPF. In our simulations, mathematics test scores in Y_0 are drawn from the normal distribution, truncated corresponding to the empirical distribution in Italy. Figure 1 displays the relationship between mathematics test scores in Y_0 and Y_1, for different values of α.

1.2 Student sorting: γ

In reality, high ability students are more likely to end up in better schools. The sorting of students on ability into better schools is captured by γ. For each student, a random number is drawn around the student's mathematics score (M_0), with standard deviation $1000(\gamma)^4$). Based on this individual number, students are assigned to equally sized schools, where the highest numbers are assigned to the best schools (measured as VA). For $\gamma = 1$, students are assigned to schools in a random manner, avoiding bias from student sorting. As γ approaches 0, students sort themselves into schools based on

Figure 1: Functional form of the EPF as a function of α.



mathematics scores in Y_0. For $\gamma = 0$, sorting is perfect: the first x spots available in the school with the highest VA are taken by the x highest performers.

Next, the leave-out-mean is calculated for each student and included in the DGP as M^p_{0i}. A student's test score in Y_1 is influenced by his peers (M^p_{0i}) following the same functional form, see (2). Depending on the value of β_p, student sorting affects predictions of M_1. In accordance with the literature on peer effects (Sacerdote, 2014), we evaluate two scenarios where peer effects are either nonexistent (β_p=0) or moderate (β_p=0.1), as "half the studies do not find evidence of peer effects in test scores [and] approximately half the studies find either modest or large effects on test scores" (Sacerdote, 2014, p.269). Note that, in the absence of peer effects, β_p=0, Y_1 mathematics test scores are still correlated to school-level average test scores, as the sorting of students groups students of high ability into high VA schools, and vice versa.

In addition to sorting on ability (test scores in Y_0), we explore an alternative scenario where low SES students (20 percent of the population) are discriminated against when they are assigned a school. In our simulation, this is done by reducing the number assigned to low SES student by 1 SD of test scores in Y_0. As a result, low SES students are three times less likely to enroll in a school that is in the top quartile in terms of school value-added. Note that SES status and test scores in Y_0 are assumed to be independent, and it does not appear directly in the education production function. Test scores in Y_1 are only affected through weaker peers and lower VA schools, as low SES students on average sort into schools with weaker peers – holding test scores constant. In Table A2, results are shown when peer effects are assumed to be zero, hence only the latter mechanism applies here. When making predictions, a dummy for low SES is included in the regression equation of the conventional approach and as an additional variable in the random forest model.

1.3 School VA: μ

The size of the school VA reflects the literature on school effects (Hattie, 2008, p.74), suggesting a small (range of 0.1 SD of M_0), or intermediate (range of 0.3 SD) effect.[9] Schools are assigned a VA randomly drawn from a normal distribution where the range reflects the assumed size of the effect. For example, when assuming an intermediate effect, the VA for each school will be drawn from a N(0, 0.3) distribution, truncated to cover a range of 0.3. We evaluate two scenarios of school value-added. First, the added value μ_s is assumed to be constant for all students within the same school, but different across schools. Second, we assume schools to only make a difference for low SES students (20 percent of the population). As such, we calculate test scores in Y_1 using two education production functions, one for each group of students. Again, when making predictions, a dummy for low SES is included in the regression equation of the conventional approach and as an additional variable in the random forest model.

2 Results

Tables A1 and A2 present the simulation results for different DGPs (averaged over B). Each table contains results for the conventional OLS approach and the RF approach advocated in this paper. Numbers displayed are differences between conventional and RF results. Hence, negative numbers in the top two panels indicate an improvement in accuracy (i.e. lower MSE), while positive numbers in the bottom two panels indicate an improvement in ranking accuracy (i.e. higher rank correlation). Values of α and γ between 0 and 1 are considered, and we allow different scenarios for the importance of μ_s in the DGP. All conventional and RF estimates are obtained by including both the individual score (M_0) and the leave-out-mean (M_0^p) as predictors of scores in Y_1 (M_1). In the first scenario (Table A1), we set $\beta_p = 0.1$ such that student sorting affects scores in Y_1 through peer effects in addition to school VA. In the second scenario (Table A2), we simulate the trivial case where peer effects do not affect individual achievement, $\beta_p = 0$. Hence, M_1 is defined by the school value-added μ_s, previous scores M_0, and measurement error ε. Under both scenarios, we can draw a similar general conclusion: The advantage of RF over conventional estimates is especially pronounced when the education production function is not *strictly* linear, and if students are not randomly assigned to schools (i.e. there is some degree of sorting on ability). If the above conditions *do* hold, jointly, it can be preferable to apply the conventional approach to estimate the school VA, and rank schools accordingly. Under all alternative parameter configurations, RF estimates provide a more accurate representation by minimizing prediction errors.

Table A3 evaluates alternative specifications under a moderate scenario ($\alpha = 0.5$ and $\gamma = 0.5$)[10]. The first two columns replicate the main simulation set-up, using these parameters. Next, results are provided when low SES students face lower chance of being assigned to better schools, independent of ability (see 1.2). The final two columns allow for heterogeneous school effects, where only low SES students are affected (see 1.3). In both additional scenarios, the random forest is better able to predict test scores in Y1, resulting in more accurate VA estimates, and school rankings.

[9] In Italy, the difference in VA between the average school in the top quartile and the average school in the bottom quartile is estimated to equal approximately 0.16 standard deviations (see 4.2).

[10] In Italy, we measure a correlation of 0.38 between Y_0 mathematics test scores and school-level average test scores. This corresponds to a level of $\gamma \approx 0.55$. Simulation results for alternative parameter combinations are available upon request.

Table A1: Prediction errors and school rankings (β_p=0.1).

Size of school VA (μ):		0.1						0.3					
Student sorting (γ):		0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
Decrease in prediction error: MSE													
Functional form (α):	0	-0.73	-0.74	-0.72	-0.76	-1.18	-1.52	-1.13	-1.13	-1.12	-1.96	-5.73	-8.90
	0.2	-71.72	-71.60	-64.41	-51.07	-49.98	-49.93	-71.53	-71.72	-64.46	-52.25	-54.50	-57.39
	0.4	-283.83	-284.37	-254.94	-202.01	-196.31	-196.35	-285.22	-283.43	-258.01	-203.41	-201.39	-203.18
	0.6	-639.26	-640.81	-574.12	-455.04	-441.93	-437.97	-637.16	-636.77	-574.49	-451.84	-446.15	-446.10
	0.8	-1140.54	-1137.24	-1019.50	-808.16	-785.03	-777.00	-1136.04	-1134.91	-1016.75	-808.29	-787.08	-785.17
	1	-1774.04	-1770.40	-1589.97	-1256.17	-1219.43	-1212.26	-1775.94	-1783.94	-1589.89	-1261.63	-1221.48	-1221.55
Decrease in VA error: MSE													
Functional form (α):	0	0.04	0.04	0.02	0.07	0.31	0.49	0.43	0.43	0.34	0.86	3.11	4.96
	0.2	-70.59	-70.29	-32.18	-0.88	-0.18	0.00	-70.14	-69.72	-31.56	-0.07	2.58	4.42
	0.4	-281.73	-281.65	-128.37	-3.69	-1.64	-1.46	-282.57	-280.24	-128.64	-2.86	1.16	2.92
	0.6	-634.73	-634.93	-290.52	-8.60	-4.00	-3.72	-634.45	-630.93	-288.29	-7.44	-1.18	0.65
	0.8	-1133.87	-1127.30	-514.81	-14.80	-7.48	-7.24	-1129.95	-1125.41	-511.57	-14.06	-4.49	-2.65
	1	-1763.56	-1754.63	-803.31	-22.94	-11.94	-11.94	-1767.12	-1768.72	-801.94	-22.61	-9.24	-7.03
Increase in Spearman rank correlation													
Functional form (α):	0	-0.26	-0.26	-0.03	0.01	-0.03	-0.06	-0.32	-0.30	-0.06	-0.01	-0.05	-0.08
	0.2	0.13	0.14	0.30	0.18	0.15	0.10	0.09	0.10	0.25	0.02	-0.02	-0.05
	0.4	0.15	0.15	0.30	0.28	0.36	0.33	0.12	0.12	0.29	0.11	0.05	0.01
	0.6	0.16	0.14	0.28	0.32	0.44	0.48	0.14	0.14	0.30	0.19	0.14	0.08
	0.8	0.16	0.17	0.28	0.35	0.51	0.54	0.14	0.16	0.29	0.24	0.21	0.17
	1	0.13	0.19	0.27	0.37	0.57	0.61	0.14	0.17	0.31	0.26	0.28	0.24
Increase in Kendall rank correlation													
Functional form (α):	0	-0.20	-0.20	-0.04	-0.01	-0.03	-0.09	-0.25	-0.24	-0.07	-0.02	-0.05	-0.11
	0.2	0.13	0.14	0.22	0.10	0.11	0.11	0.09	0.09	0.16	-0.02	-0.02	-0.07
	0.4	0.15	0.15	0.23	0.18	0.26	0.30	0.13	0.12	0.20	0.04	0.04	0.01
	0.6	0.16	0.15	0.22	0.22	0.32	0.41	0.15	0.14	0.22	0.10	0.10	0.09
	0.8	0.16	0.17	0.22	0.23	0.37	0.45	0.15	0.16	0.22	0.14	0.16	0.16
	1	0.14	0.18	0.21	0.25	0.41	0.50	0.15	0.17	0.23	0.16	0.21	0.23

Table A2: Prediction errors and school rankings (β_p=0).

Size of school VA (μ):		0.1						0.3					
Student sorting (γ):		0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
Decrease in prediction error: MSE													
Functional form (α):	0	-0.73	-0.74	-0.72	-0.77	-1.16	-1.52	-1.14	-1.17	-1.15	-2.03	-5.80	-9.01
	0.2	-28.88	-28.72	-29.06	-29.18	-29.43	-29.91	-29.08	-28.87	-29.41	-30.36	-34.35	-37.31
	0.4	-114.11	-113.92	-113.89	-114.11	-114.88	-114.67	-113.41	-112.88	-114.99	-115.26	-120.02	-122.34
	0.6	-254.76	-255.50	-255.22	-256.06	-256.27	-256.76	-254.86	-253.93	-256.31	-256.20	-261.07	-264.20
	0.8	-454.37	-453.67	-450.96	-456.54	-455.51	-453.01	-448.82	-451.65	-453.99	-455.44	-460.93	-462.12
	1	-706.20	-710.69	-709.07	-713.04	-708.52	-711.93	-706.28	-709.87	-708.53	-709.54	-716.02	-716.29
Decrease in VA error: MSE													
Functional form (α):	0	0.04	0.04	0.03	0.08	0.30	0.48	0.45	0.48	0.40	0.93	3.13	5.00
	0.2	-28.15	-27.84	-12.43	-0.36	0.02	0.20	-27.76	-27.54	-12.08	0.45	2.94	4.71
	0.4	-113.05	-112.25	-50.01	-1.63	-0.81	-0.63	-112.41	-111.31	-49.70	-0.75	2.10	3.85
	0.6	-252.87	-252.76	-112.10	-3.86	-2.19	-2.05	-253.90	-252.17	-111.14	-2.98	0.77	2.67
	0.8	-451.24	-449.15	-197.39	-6.80	-4.09	-3.90	-448.23	-447.32	-198.86	-6.00	-1.00	0.77
	1	-702.38	-703.81	-311.89	-10.51	-6.72	-6.35	-703.15	-704.19	-311.04	-9.90	-3.96	-1.70
Increase in Spearman rank correlation													
Functional form (α):	0	-0.28	-0.28	-0.14	-0.02	-0.03	-0.05	-0.30	-0.30	-0.11	-0.02	-0.05	-0.07
	0.2	0.15	0.17	0.21	0.10	0.09	0.05	0.11	0.11	0.16	-0.01	-0.04	-0.06
	0.4	0.17	0.18	0.24	0.20	0.26	0.22	0.15	0.13	0.24	0.05	0.01	-0.03
	0.6	0.18	0.21	0.25	0.27	0.37	0.36	0.17	0.18	0.26	0.11	0.06	0.02
	0.8	0.18	0.20	0.27	0.30	0.45	0.45	0.18	0.19	0.29	0.16	0.13	0.08
	1	0.20	0.19	0.29	0.31	0.50	0.53	0.20	0.18	0.31	0.21	0.20	0.14
Increase in Kendall rank correlation													
Functional form (α):	0	-0.21	-0.21	-0.12	-0.03	-0.03	-0.08	-0.23	-0.23	-0.10	-0.04	-0.05	-0.11
	0.2	0.16	0.17	0.16	0.05	0.06	0.05	0.10	0.10	0.08	-0.03	-0.04	-0.09
	0.4	0.18	0.18	0.19	0.12	0.19	0.22	0.15	0.14	0.17	0.00	0.00	-0.03
	0.6	0.19	0.20	0.20	0.17	0.27	0.32	0.17	0.18	0.18	0.05	0.04	0.02
	0.8	0.19	0.20	0.22	0.19	0.32	0.39	0.18	0.19	0.22	0.09	0.09	0.08
	1	0.20	0.19	0.23	0.20	0.36	0.44	0.20	0.19	0.24	0.12	0.14	0.14

Table A3: Alternative specifications ($\alpha = 0.5$ and $\gamma = 0.5$).

	Sorting on ability		Sorting on ability and demographics		Heterogeneous effects	
Size of school VA (μ):	0.1	0.3	0.1	0.3	0.1	0.3
(β_p=0.1)						
MSE	-342.59	-342.49	-145.72	-144.21	-148.60	-147.00
MSE VA	-33.56	-33.33	-17.90	-33.47	-18.00	-33.58
SP	0.27	0.21	0.95	0.87	0.98	0.97
KEN	0.18	0.11	0.84	0.78	0.86	0.86
(β_p=0)						
MSE	-178.77	-179.25	-86.66	-86.36	-88.02	-88.80
MSE VA	-13.05	-12.50	-9.09	-20.19	-9.51	-18.94
SP	0.19	0.12	0.94	0.84	0.98	0.98
KEN	0.12	0.05	0.84	0.76	0.86	0.86

Notes: MSE and MSE VA denote the error reduction in test score predictions and VA estimates, respectively, relative to conventional estimates. A negative number implies an improvement in accuracy by using the RF model. SP and KEN denote Spearman and Kendall correlation coefficients, respectively, relative to conventional estimates. A positive number implies a more accurate ranking of schools when using the RF model.

Parameters used in Monte Carlo simulations:

- α: Nonlinearity of education production function.

  This parameter captures the relationship between test scores in Year 0 (peers' and students' own scores) and test scores in Year 1. For α=0, this relationship is strictly linear, whereas for α=1, the functional form corresponds to the polynomial function specified in section 1.1.

- γ: Degree of sorting of students into schools.

  When γ equals 1, all students are equally likely to be assigned the best school. The rank order correlation between a student's own test scores and the rank of his or her school is close to 0. When γ equals 0, students are assigned to schools following a deterministic rule: the school with the highest VA will consist of the 100 best students etc. In this other extreme setting, the rank order correlation between a student's own test scores and the rank of his or her school is 1.

- μ_s: The added value of school s.

  Drawn from the normal distribution around 0, with a range equal to either 0.1 or 0.3 standard deviations of Year 0 test scores.

- β_p: Peer effects.

  This parameter indicates the importance of peers' test scores on students' own scores. Peer effects are measured as the leave-out-mean for student i in school s (M_i^p), and the functional form linking M_i^p to test scores in Year 1 is determined by α.