Schiltz, Fritz; Masci, Chiara; Agasisti, Tommaso; Horn, Dániel

**Working Paper**

# Using machine learning to model interaction effects in education: A graphical approach

Budapest Working Papers on the Labour Market, No. BWP - 2017/4

**Provided in Cooperation with:**
Institute of Economics, Centre for Economic and Regional Studies, Hungarian Academy of Sciences

*Suggested Citation:* Schiltz, Fritz; Masci, Chiara; Agasisti, Tommaso; Horn, Dániel (2017) : Using machine learning to model interaction effects in education: A graphical approach, Budapest Working Papers on the Labour Market, No. BWP - 2017/4, ISBN 978-615-5594-99-1, Hungarian Academy of Sciences, Institute of Economics, Centre for Economic and Regional Studies, Budapest

This Version is available at:
https://hdl.handle.net/10419/200355


**BUDAPEST WORKING PAPERS ON THE LABOUR MARKET**

**BWP – 2017/4**

# Using Machine Learning to Model Interaction Effects in Education: A Graphical Approach

FRITZ SCHILTZ, CHIARA MASCI, TOMMASO AGASISTI

AND DANIEL HORN

BWP 2017/4

INSTITUTE OF ECONOMICS, CENTRE FOR ECONOMIC AND REGIONAL STUDIES
HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST, 2017

Budapest Working Papers on the Labour Market

BWP – 2017/4

Institute of Economics, Centre for Economic and Regional Studies,
Hungarian Academy of Sciences

# Using Machine Learning To Model Interaction Effects In Education: A Graphical Approach

Authors:

Fritz Schiltz
researcher
Leuven Economics of Education Research, University of Leuven, Belgium
E-mail: fritz.schiltz@kuleuven.be

Chiara Masci
PhD student
Modelling and Scientific Computing, Department of Mathematics,
Politecnico di Milano, Italy
E-mail: chiara.masci@polimi.it

Tommaso Agasisti
associate professor
Department of Management, Economics and Industrial Engineering,
Politecnico di Milano, Italy
E-mail: tommaso.agasisti@polimi.it

Daniel Horn
senior research fellow
Centre for Economic and Regional Studies, Hungarian Academy of Sciences, Hungary
E-mail: horn.daniel@krtk.mta.hu

June 2017

ISBN 978-615-5594-99-1
ISSN 1785 3788

# Using Machine Learning To Model Interaction Effects In Education: A Graphical Approach

Fritz Schiltz, Chiara Masci, Tommaso Agasisti and Daniel Horn

## Abstract

Educational systems can be characterized by a complex structure: students, classes and teachers, schools and principals, and providers of education. The added value of schools is likely influenced by all these levels and, especially, by interactions between them. We illustrate the ability of Machine Learning (ML) methods (Regression Trees, Random Forests and Boosting) to model this complex 'education production function' using Hungarian data. We find that, in contrast to ML methods, classical regression approaches fail to identify relevant nonlinear interactions such as the role of school principals to accommodate district size policies. We visualize nonlinear interaction effects in a way that can be easily interpreted.

Keywords: machine learning, education production function, interaction effects, non-linear effects

JEL codes: C5, C18, C49, I21, H75

## Acknowledgement:

We are grateful to Geraint Johnes, Daniel Santin and other seminar participants in Milan, Lisbon, Leuven, and Budapest, for many useful comments and remarks and to Anna Maria Paganoni for the statistical support during the analysis. The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 691676 (EdEN).

# Gépi tanulás használata az interakciós hatások modellezésére az oktatásban: egy grafikus megközelítés

Fritz Schiltz, Chiara Masci, Tommaso Agasisti, Horn Dániel

## Összefoglaló

Az oktatási rendszer több egymásba ágyazott szereplőből áll: diákok, osztályok és tanárok, iskolák és igazgatók, és iskolafenntartók. Az iskola hozzáadott értékét az összes szint és ezek interakciói is befolyásolhatják. A tanulmány a „gépi tanulás" (Machine Learning – ML) hasznosságát illusztrálja az oktatási termelési függvény becslés esetében, magyar adatokon. Bemutatjuk, hogy szemben a felhasznált ML módszerekkel (Regression Trees, Random Forests és Boosting), a klasszikus regressziós eljárások nem azonosítják jól a nem lineáris összefüggéseket. Egy ilyen példa az iskolaigazgatók szerepe a különböző méretű iskolakörzetekben. A nem lineáris összefüggéseket vizuális módszerekkel mutatjuk be, a könnyebb érthetőség miatt.

Tárgyszavak: gépi tanulás, oktatási termelési függvény, interakciós hatások, nem lineáris hatások

JEL kódok: C5, C18, C49, I21, H75

## 1. INTRODUCTION

Machine learning (ML) methods are increasingly being used in different fields of economics, for example in predicting worker productivity (Chalfin et al., 2016), poverty alleviation (Blumenstock, 2016), and have been put forward as a useful tool in the econometric toolbox (Mullainathan & Spiess, 2017). However, applications to education only received little attention (Vanthienen & De Witte, 2017). In education economics, student achievement can be seen as the outcome of a production process characterized by many interactions between different stakeholders. Analogous to the production function in social policy applications, Y=f(L,K), this process has been modelled as the education production function.1 Econometricians can offer two contributions to the study of this EPF and its components. The first is causal inference for a specific component and the estimation of its net contribution to student achievement. For example, Gerritsen, Plug, & Webbink (2017) provide evidence on the importance of teacher quality and experience for student achievement using data on twin pairs. The second, and the focus of this paper, is to provide insights into the relative importance of, and interaction between, those components.2 This is particularly interesting in the context of education. Many decisions are taken by different actors at different levels of operations (school, class, provider), and as a direct consequence, researchers and policymakers that do not acknowledge these interactions will over- or underestimate the anticipated impact of education policies. This paper introduces machine learning methods (or rather "supervised" machine learning), and illustrates how this approach can be useful when modelling an (education) production function where stakeholders interact within and between different levels.

In the diverse educational landscape, it could for example be the case that some actions of school principals and other-level decision makers can be effective in specific school contexts, and have no (or negative) effect in others. Moreover, context-dependencies have been put forward as an explanation of conflicting results in the literature (Burgess, 2016). Incorporating the "context" in the estimation method will

[1] For a description of the EPF approach, see Hanushek & Woessmann (2010).
[2] Note that recently, Athey & Imbens (2017) outlined possibilities for ML methods to combine both contributions: supplementary analysis using ML in quasi-experimental settings. Although this is not the aim of this study, our approach can be easily extended to these types of settings and datasets.

likely produce more consistent results[3]. A common approach followed by economists and other social scientists (in education and other fields) is to include multiplicative interactions. However, the execution of these models is often flawed due to the lack of conditional hypotheses and interpretation errors (Brambor, Clark, & Golder, 2006). A conditional hypothesis such as "an increase in X is associated with an increase in Y when condition Z holds" implies the need for an *a priori* specification of this interaction effect. If the interaction term is not specified, it will not be estimated in a standard regression approach. In other words, some preconception is needed with respect to functional form of the EPF. In addition, interaction effects are often included linearly.

Although this paper is not the first to apply machine learning using education data[4], it is the first, to the best of our knowledge, to apply machine learning methods to model and visualize interaction effects in education. We use machine learning as a piece of the theory. We acknowledge the existence of a relationship between different stakeholders in the 'production' of student achievement, but we choose to derive these relationships from the data instead of imposing a functional form (James, Witten, Hastie, & Tibshirani, 2013). While statistical approaches to model-fitting start by assuming a functional form and then estimate the parameters from the data, the ML approach uses an algorithm to learn the relationship between the outcome and its predictors, avoiding to impose any functional form. Moreover, an ML approach assumes that the data-generating process is complex and unknown and it tries to learn the dominant patterns between the predictors and outcome variables. The model identifies important interactions thanks to its tree-structure, without requiring the researcher to have any preconceptions on this matter.

We illustrate the approach using Hungarian data. Our data covers the 2008-2010 period and includes variables with respect to all organizational levels in primary education (student, class, school, and education provider). The data allows estimating a value-added approach since the *same* students are followed over time. Our findings indicate that classical regression approaches are to some extent able to capture the important variables but fail to identify interesting nonlinear interactions between and

[3] The 'educational context' can be interpreted broadly. For example, when estimating the optimal class size, the schools where classes are based can be considered the context. As Loveless & Hess (2007) note: "*Small classes tend to be clustered in small schools, and average class size is larger in large as compared to small schools. In this way school size effects might 'work' indirectly through smaller classes, as intermediary conditions*."

[4] Thomas & Galambos (2004) apply regression and decision trees to investigate how students' characteristics and experiences affect satisfaction. Ma (2005) follows a two-stage approach in which in the first stage, he estimates the rate of growth in mathematics achievements of each student, by means of hierarchical linear model (HLM), while in the second stage he applies classification and regression trees (CART) to relate them to students' characteristics. Cortez & Silva (2008) apply Data Mining (DM) methods as regression trees and random forest to relate Portuguese students' scores in mathematics and reading to students' characteristics.

within levels. We introduce machine learning methods that capture these nonlinearities and visualize them in order to improve interpretability.

The remainder of the paper is organized as follows. Section 2 introduces our empirical strategy. A brief overview of the Hungarian data is presented in section 3, followed by the discussion of our graphical results in section 4. Section 5 concludes, and suggests outlines for further research in econometrics.

## 2. EMPIRICAL STRATEGY

### 2.1 THE ADDED VALUE OF SCHOOLS

Explanatory variables at different levels are generally linked to student outcomes using an educational production function (EPF) approach. The common approach in the economic literature is to estimate value-added models. By including measures of prior achievement, these models focus on the *change* in student achievement over the specified time period (i.e. from prior achievement to estimated achievement). To overcome the assumption that the functional form of achievement is linear and additively separable (Todd & Wolpin, 2003), we construct a measure of school value-added using *Data Envelopment Analysis (DEA)*. DEA constructs a frontier based on the data and allows comparison of schools with their reference school, given inputs, without any assumption on the functional form.[5] Apart from its nonparametric structure, DEA offers two additional advantages. First, the ability of DEA to handle multiple inputs and outputs allows us to leave out socio-economic background and prior achievement as right-hand-side variables of the EPF model and include them as inputs in the linear programming problem. It is well established, and intuitive, that these two variables are the most important predictors of student achievement (e.g. Haveman & Wolfe, 1995). If instead, prior achievement and socio-economic background are included when building regression trees to model the EPF (see below), the relative importance in terms of explained variability of other explanatory variables would become very moderate. As we are mainly interested in studying the importance of managerial inputs (e.g. class, school and provider size, and principal characteristics), it is better to include the socio-economic background and prior achievement in the DEA specification (see Online Appendix A). Second, conceptualizing the added value of schools in an efficiency framework allows a clear interpretation in both classical regression and ML models. That

[5] This reference school is not the average school, but a school that is situated on the 'efficiency frontier'. The frontier for a given school consists of schools attaching the same weights to the inputs (prior achievement and socio-economic background).

is, the added value of a school can be defined as its ability to transform a set of inputs into outputs. For example, a school with an efficiency score, or added value, of 80%, can be evaluated as follows: If this school would be as efficient as its reference group, given prior performance and socio-economic background (inputs), it could improve the performance (output) of its students by 20%. We are aware that there are other methods to obtain measures of school value-added, but these do not offer the aforementioned advantages when machine learning methods are used.

An important condition when applying DEA, and a potential flaw, is the accuracy of the data. Measurement errors shift the frontier and hence bias the obtained efficiency scores (or, value-added measures). Therefore, we generate bias-corrected estimates of the added value of schools, following the approach of Simar & Wilson (2007), and set the number of bootstrap replications at 2000. We specify an output-orientation and allow for variable returns to scale.[6]

### 2.2 MODELLING THE EPF USING MACHINE LEARNING

The aim of the second stage in our analysis is to identify variables that are associated with a high added value of schools. We apply regression trees and random forests (see James, Witten, Hastie, & Tibshirani, 2013) to identify the importance of, and relationship between, components of the EPF.

One issue when modelling the EPF is the selection of its components. In this paper we are interested in managerial variables - that can be influenced by policy-makers - and how they (inter)relate to the added value of schools.[7] In particular, class, school, and provider size, and principal characteristics. All these variables have been studied extensively, and although this literature remains unsettled, a recurring notion is that results are context-dependent in the multi-governance structure of education (Burns & Köster, 2016). For example, Bloom et al. (2015) point at the importance of including other levels of governance in education when estimating the impact of school principals and management using a (linear) production function. They find that principal added value is highly subject to provider strategy and accountability. As a result, leaving out the provider level biases coefficients on principal variables.

Apart from the issue of selecting inputs to be included in the EPF, assumptions need to be made on its functional form. As we introduced before, in the (education) economics literature, a linear functional form is frequently imposed on the EPF. Regression trees

[6] The development of the linear programming problem is further elaborated in Online Appendix A.

[7] Note that prior achievement and socio-economic background are already included when generating the school added value (see 2.1.).

have a very different flavor compared to these more classical regression approaches. In particular, a linear regression model assumes the following functional form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \quad (1)$$

with $X_j$ a matrix of $p$ predictors, or components of the EPF. Regression trees assume a model of the form:

$$f(X) = \sum_{m=1}^{M} c_m I_{(X \in R_m)} \quad (2)$$

where $R_1$,.., $R_M$ represent a partition of the predictor space. Determining which model is more appropriate depends on the problem: if the relationship between the components and the outcome is well approximated by a linear model, then an approach such as linear regression will likely work well, and will outperform a method such as a regression tree that does not exploit this linear structure (Varian, 2014). If instead there is a highly non-linear and complex relationship between the components and the outcome, then decision trees may outperform classical approaches.

Given an outcome variable and a set of predictors, *tree-based methods*, for regression or classification, involve a segmentation of the predictor space into a number of regions. In order to make a prediction for a given observation, we typically use the mean or the mode of the observations in the region to which it belongs. Roughly speaking, building a regression tree involves two steps:

1. We divide the predictor space - that is, the set of possible values for components $X_1$, $X_2$,.., $X_p$ into J distinct and non-overlapping regions, $R_1$, $R_2$,.., $R_j$. For simplicity, we consider these regions as high-dimensional rectangles (or boxes);
2. For every observation that falls into the region $R_j$, we make the same prediction, which is simply the mean of the response values for the observations in $R_j$.

The regions are chosen in order to minimize the Residual Sum of Squares (RSS) where $\hat{y}_{R_j}$ is the mean of the observations within the j-th box and $y_{ij}$ is the i-th observation within the j-th box:

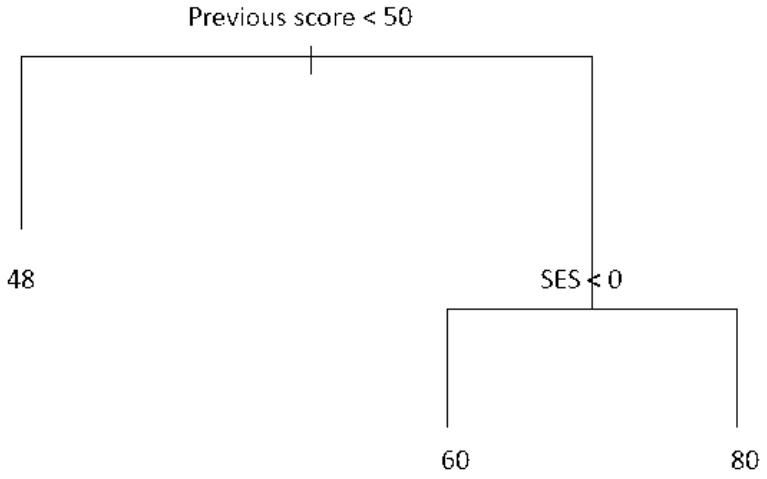
$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_{ij} - \hat{y}_{R_j})^2 \quad (3)$$

In order to explain the idea of regression trees, consider the following example. Imagine we want to regress student test scores (a continuous variable taking on values between 0 and 100), on students' previous scores, their socio-economical index (SES – a continuous variable having mean 0 and variance 1) and their gender (M/F). Then the regression tree we obtain can be displayed as in Figure. The algorithm is able to identify

the most important variables in explaining the response (previous score and SES) and, in particular, the threshold values (previous score = 50 in the first split, SES = 0 in the second one) that, at each split, are able to divide the population in two subgroups, minimizing the variability within each group. Looking at the tree below, we can conclude that the only two variables that matter are the previous students' scores and the SES - in this hypothetical example. This implies that gender is not able to catch any variability between student test scores – otherwise, it would be included in the regression tree. Moreover, the most important variable (on top of the tree) is the previous test score. Hence, when estimating a student's score, we read the tree in this way: if the previous score of a student is less than 50, then the estimated student score is 48, while if the previous score of a student is bigger than 50, it depends on the student SES: if the student SES is higher than 0, the expected student score is 80, while if it is less than 0, the expected score is 60. The values on the leaves of the trees are obtained by averaging the mean of the scores of students that fall in the same 'leaf'.

*Figure 1*

**A simple regression tree**




Note: Outcome variable is student test score (continuous, ranging from 0 to 100), and the two predictors selected by the regression tree are previous score (continuous, ranging from 0 to 100) and socio-economic index - SES (continuous with mean = 0 and standard deviation = 1).

There are four main advantages of regression trees in an education context. First, trees can be displayed graphically and are very easily interpretable, even in dialogues with non-technical experts and decision makers. Second, estimating the EPF using regression trees does not force any type of relationship between outcome variables and components of the EPF. Third, regression trees can easily handle qualitative predictors

without the need to create dummy variables. Fourth, interactions are allowed between variables, by explicitly modelling how groups of inputs are associated with different levels of output production.

This last point is worth stressing as it is the major strength of regression trees in the context of education. As displayed in Figure 1, there is an interaction between (and within) different levels in the educational system. Indeed, while in a linear regression interactions between variables need to be specified *a priori*, trees investigate which are the interactions that matter, without forcing any linearity (or other functional forms) in their relationship.

Nevertheless, and despite these advantages, regression trees are also characterized by some disadvantages: they generally suffer from high variance and are sensitive to outliers. However, by aggregating many decision trees, using methods like *bagging*, *random forests* and *boosting*, the predictive performance of trees can be substantially improved (see again James et al., 2013). The following three methods use trees as building blocks in order to construct more powerful models.

**Bagging**. A natural way to reduce the variance and hence increase the predictive power of a statistical learning method is to take bootstrapped samples from the population, build a separate prediction model using each training set, and average the resulting predictions. In other words, *bagging* calculates $\hat{f}_1(x), \hat{f}_2(x), \ldots, \hat{f}_B(x)$ using B separate bootstrapped training sets, and averages them in order to obtain a single low-variance statistical learning model:

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x) \quad (4)$$

**Random forests**. This second approach provides an improvement over *bagging* thanks to a small tweak that de-correlates the trees. As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of *m* predictors is chosen as split candidates from the full set of p predictors. A fresh sample of *m* predictors is taken at each split. This approach allows all the predictors to be taken into account and to differentiate the trees that will be averaged.

**Boosting**. Lastly, *boosting* works like *bagging*, except the fact that the trees are grown *sequentially*: Each tree is grown using information from previously grown trees. *Boosting* does not involve bootstrap sampling; instead each tree is fit on a modified version of the original data set. The idea behind this procedure is that, unlike fitting a single large decision tree to the data, which amounts to fitting the data hard and potentially overfitting (Varian, 2014), the boosting approach instead learns slowly. Given the current model, a decision tree is fitted to the residuals of the model, rather than the

outcome variable. Note that in boosting, unlike in bagging, the construction of each tree depends strongly on the trees that have already been grown. This new decision tree is then added into the fitted function in order to update the residuals (see James et al., 2013).

These three methods considerably increase the (predictive) power of our model and make it more robust to the presence of outliers. However, a price needs to be paid in terms of interpretability, since applying them instead of a simple regression tree, it is no longer possible to graphically display the final tree (as in Figure 1). Nonetheless, it is possible to analyze the results by looking at the importance of the predictors, and to display the relationships between explanatory and outcome variables in an intuitive way (see section 4). In other words, we have information on the percentage of variability explained by the model, the *importance* of each predictor, measured as the ability of each predictor to reduce the node purity (measured by Residual Sum of Square), and the partial influence of each predictor or the combined influence of two predictors on the outcome variable.

## 3. DATA

Our dataset was constructed by integrating information on student, class, and school characteristics from the National Assessment of Basic Competencies (NABC) of Hungary. The NABC covers all students in primary schools in Hungary.[8] It is a standard based assessment for mathematics and reading that follows the model of the Programme for International Student Assessment (PISA), however, it is conducted annually in May. Students are tested before grade 6 (age 12) and before graduation from primary school, in grade 8 (age 14). In addition to mathematics and literacy test scores, which are common to education datasets, our database contains extensive information on the student background, principals, and the school. This study uses data on the 6$^{th}$ grade 2008 cohort, graduating primary schools in 2010. Our main analysis is performed on school level NABC data. Because of missing values with respect to the outcome and explanatory variables, the final dataset contains 2122 schools. All variables used in subsequent analyses are presented in Table 1a-1b. A further description of the data used here can be found in Kertesi & Kezdi (2011).

In order to obtain a measure of the added value of schools, input and output variables were averaged for every school. The variable indicating the status of the students is a z-standardized socio-economic status (SES) index, with 0 mean 1 standard

[8] A brief introduction to the Hungarian education system is presented in Online Appendix B.

deviation. This index follows the economic-social and cultural status (ESCS) index of the OECD PISA studies.

*Table 1a*

**Descriptive statistics**

| Variables | N | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| School size | 2,122 | 21 | 6 | 3 | 37 |
| Class size (school average) | 2,122 | 292 | 187 | 0 | 2,195 |
| % Roma students | 2,122 | 16.46 | 22.43 | 0 | 100 |
| Number of computers# | 2,122 | 17.47 | 6.617 | 0 | 80 |
| Teacher with training (%) | 2,122 | 32.17 | 28.49 | 0 | 100 |
| Experience principal (years) | 2,122 | 8.238 | 6.532 | 0 | 55 |
| Age principal (years) | 2,122 | 56.40 | 6.507 | 31 | 75 |
| Principal satisfaction (%)* | 2,122 | 71.10 | 20.87 | 0 | 100 |
| Provider size | 2,122 | 7.48 | 12.06 | 1 | 97 |
| School added value, Math | 2,122 | 0.53 | 0.11 | 0.13 | 0.89 |
| School added value, Reading | 2,122 | 0.71 | 0.11 | 0.18 | 0.99 |

*Table 1b*

**Descriptive statistics (continued)**

| **Categorical variables (N=2122)** | | | | | | | |
|---|---|---|---|---|---|---|---|
| School location | Budapest | City | County Centre | Village <2k | Village (2k-5k) | Village >5k | |
| | 11% | 28% | 14% | 29% | 16% | 2% | |
| Region | Central H | Central Td | Northern GP | Northern H | Southern GP | Southern Td | Western Td |
| | 22% | 12% | 16% | 16% | 13% | 9% | 12% |
| Education Provider | Settlement / District government | Ecclesiastical | Private | Other government | Other | | |
| | 85% | 7% | 2% | 1% | 5% | | |

Note: School added value (reading and mathematics) are reported before normalization. #: 'Computers available' measures the total number of computers as counted in the dedicated computer class. *: Principal satisfaction as a percentage, representing the answer to the question "If you were to assigned to another school, what percentage of the current teaching staff would you take with you to your new place?"
GP = Great Plain, Td = Transdanubia, H = Hungary

To explore the importance of components in the EPF and the interactions that determine its shape, we add a set of controls and explanatory variables frequently studied in the literature (Burgess, 2016). To illustrate the ability of machine learning

methods to model the EPF, we include class, school, and provider size.[9] In the decentralized Hungarian system, the type and size of education providers varies widely from very small local government providers with only one school to large centralized networks of church schools. Also, the number of computers in the dedicated computer class, as a proxy for school resources, and the percentage of teachers receiving additional training are included as components of the education production function.

Complementing the administrative data, we also include several variables from a questionnaire in NABC completed by the school principal. These additional variables can be used to describe the organisational setting of the schools. The school level questionnaire includes variables such as principal experience, age, and satisfaction[10]. Also, the perceived ratio of Roma students is indicated by the principal of each school[11]. Even though it seems that the individual performance of Roma students does not differ significantly from other (non-Roma) students, once socio-economic background is accounted for (see Kertesi & Kezdi, 2011), the inherent discriminatory tendencies in the Hungarian society might cause some families (or even teachers) to refrain from enrolling in schools where large Roma ratios are present. On average, 16% of students are considered to be from Roma origin. Because of an increasingly segregated educational system where so called 'Roma schools' are being ghettoized (Kertesi & Kezdi, 2011), the median value is much lower at 8%. Finally, to capture geographical discrepancies in schools' performance, we include both regional and school location categorical variables (e.g. village, city etc.).

## 4. RESULTS

In this section we graphically display the obtained results. In comparison to classical approaches, no coefficients are obtained using the methods presented in section 2. In order to allow comparison, we included the output of OLS regressions in Table 2.

Added value of schools. As can be seen from the histograms in Figure 2, the added value of schools is higher for reading (0.71 vs 0.53). This can be interpreted as follows; on average, a school can improve their students' reading achievement by 28% if it were

[9] Since our analysis is at school level, 'class size' is measured as the average class size in a school. Provider size is measured as the number of schools under supervision of the same provider of education.

[10] All principals were asked "If you were to assigned to another school, what percentage of the current teaching staff would you take with you to your new place?".

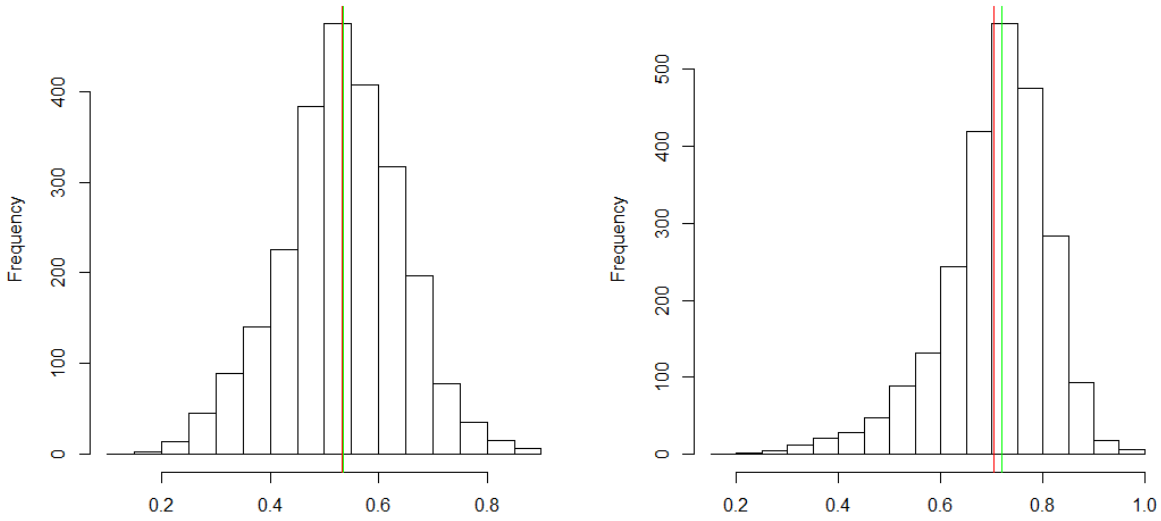
[11] This question is included in the questionnaire to overcome the stigma in Hungary with respect to Roma self-reporting. The question is the following: "In your opinion, what is the percentage ratio on your location of those among the primary school students who can be characterized by the following features? [...] Of Roma origin ... ?"

to perform as well as its reference school. Looking at the min and max values, these range from 0.13 to 0.89 and from 0.18 to 0.99, for math and reading, respectively. The variation is around 0.11 for both math and reading with the latter distribution slightly skewed to the left (see Figure 2).

In the following sections, we only present the results that were obtained using the added value of schools in terms of mathematics as the outcome variable. Also in the partial plots, joint plots, and the linear model, we restricted the displayed results to the added value of schools in mathematics. The partial and joint plots are similar in reading compared to those in mathematics. Hence, presenting all results twice will unnecessarily lengthen this paper, which aims to illustrate the intuitive graphical presentation of complex tree-based methods.

*Figure 2*

**The distribution of school added value for mathematics (left) and reading (right)**



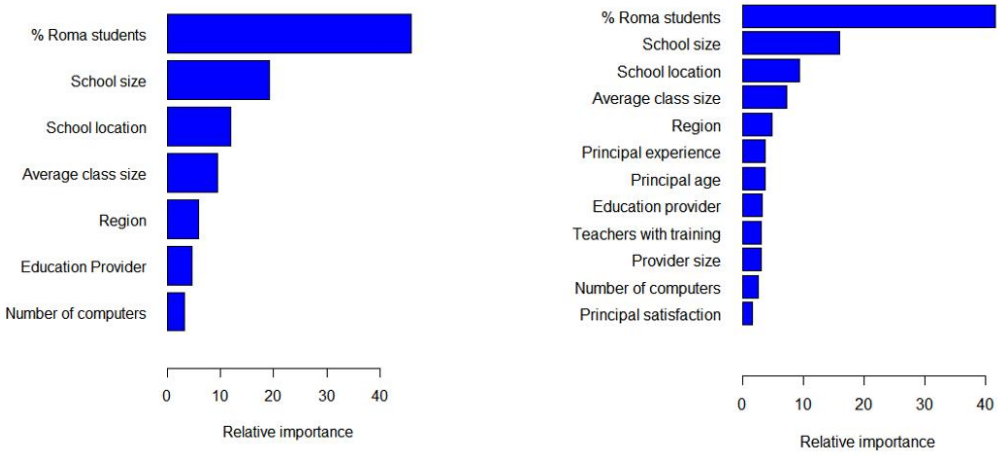

**Variance-Importance plots**. Next, to explain the variation in the added value of schools, we combine boosting and Random Forests to construct a 'forest' made up of 2000 trees. We choose $m = \sqrt{p}$ ; that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors. The outcome of the Boosting model[12] is displayed in Figure 3 where the relative importances ("Node Purity", see Methdology) of explanatory variables (both continuous and categorical) are ranked for the base model and the final model[13]. In here, we can see that

[12] Considering 3000 trees, interaction depth equal to 4, shrinkage parameter to 0.001, and 70% of the data used as training dataset.
[13] Note that the ranking of importance obtained by random forests and boosting is the same.

common variables such as average class size, school location[14], and especially school size are important. In both models, the ratio of Roma students in a school is selected as the variable most strongly related to the added value of schools.

*Figure 3*

**VI plots of base model (left) and final model (right) for mathematics**




Note: base and final models differ in the predictors involved. Base model uses 7 predictors (the ones appearing on the left panel of the figure), while the final model uses 12 predictors (the ones appearing on the right panel of the figure). It is worth to notice that Random Forest identifies the same important variables.

This confirms what we expected, considering the social stigma with respect to the Roma ethnic group (see Section 3). At the same time, this finding corroborates previous evidence about the role of the composition of student population (i.e. their socio-economic background) as a key input for educational production (Haveman & Wolfe, 1995).

When we compare the left and right panel of Figure 3, we can state that principal characteristics are closely related to the added value of schools. Also, once these variables are included, along with provider size (number of schools per provider) and the percentage of teachers receiving training, we conclude that the relative importance of class and school size diminishes. It might be, for example, that more experienced principals are in charge of larger schools, mitigating the explained variance in student performance by schools size, once principal experience is accounted for. Also, teachers having received more training might be allocated to relatively large classes. Once teacher training is accounted for, the relative importance should then reduce. Without including these characteristics, one might overstate the importance of these scale variables.

[14] 'School location' is a categorical variable indicating the geographical area – the 'site' – where a school operates. Categories include Budapest, city (not Budapest), county center, and villages (by size).

Formally, the final model outperforms, the base model in terms of (pseudo) $R^2$ (24.91 vs 18.7) and MSE (0.01007 vs 0.01031). The relative importance of variables in the education production functions roughly corresponds to the size of coefficients obtained from OLS regression and presented in Table 2 (Model 0 and Model 1). Clearly, the ratio of Roma students in a school has the strongest association with the added value of schools, followed by class size, school size and location, and principal characteristics.

Now that we know the relative imporance of the variables in our dataset, we want to know *how* these variables affect the added value of schools. In order to determine the direction and the shape of the effects, we will need to generate partial plots.[15]

**Partial plots** display the partial influence of a predictor on the outcome, averaging out the influence of all other predictors included in the model. In other words, our model isolates the net effect. The graphical approach can be compared to plotting the coefficient of a (linear) regression, without assuming this coefficient is constant (or varies at given rate) across values of a specified variable[16]. Figure 4 displays partial plots for the four most important variables, as indicated in Figure 3 (i.e. % Roma students, school size, school location, and class size). OLS regression results are included in Table 2, Model 1.

From this OLS regression we conclude that there is a significant correlation between class- and school size, and the added value of schools in Hungary. This is roughly reflected in the partial plots for class- and school size. However, if we have a closer look, we can see that the school size slope is only positive up to 400-500 students. Once this threshold is surpassed, larger schools do not appear to have a large added value. This result could indicate a saturation point of school size advantages, although OLS does not allow us to identify this levelling off of the school added value. If we were to plot the coefficient obtained in Table 2, it would suggest an ever increasing added value of increasing schools, which does not appear to hold once more flexibility in the EPF is allowed[17]. Hence, when assuming a linear trend a lot of insightful information gets lost, motivating the use of more flexible models like Random Forests and Boosting.

On the other hand, the percentage of Roma students seems to be negatively related to the school added value. Also in a simple linear regression model, we find that the share

[15] All figures in this section (partial and joint) are based on the results obtained from the final model, including all explanatory variables (and their possible interactions). For the sake of brevity, we only included a selection of graphs, while the remainder are available upon request from the authors.

[16] In this respect, partial plots obtained using Boosting can be seen as an alternative to graphical representations of quantile regressions. As we will display in the following paragraph, Boosting models offer the advantage to visualize joint partial plots.

[17] Note that the explanatory power of Boosting is actually above the $R^2$ (24.9 vs 21.9) we obtain when running an OLS regression (see Model 3 in Table 2). This presence of model improvement, as measured by $R^2$, might be due to the actual relationships not being linear (section 4.2 and Varian, 2014).

of Roma students in a school is significantly related to lower added value of schools (Table 2, Model 1). Shifting our attention to the location of the school, we conclude that schools based in Budapest and county centers outperform schools in cities and villages. Further disentangling villages by population size, we find that schools in small villages (<2k inhabitants) and relatively large villages (>5k inhabitants) outperform those in medium-sized villages in Hungary.

*Table 2*

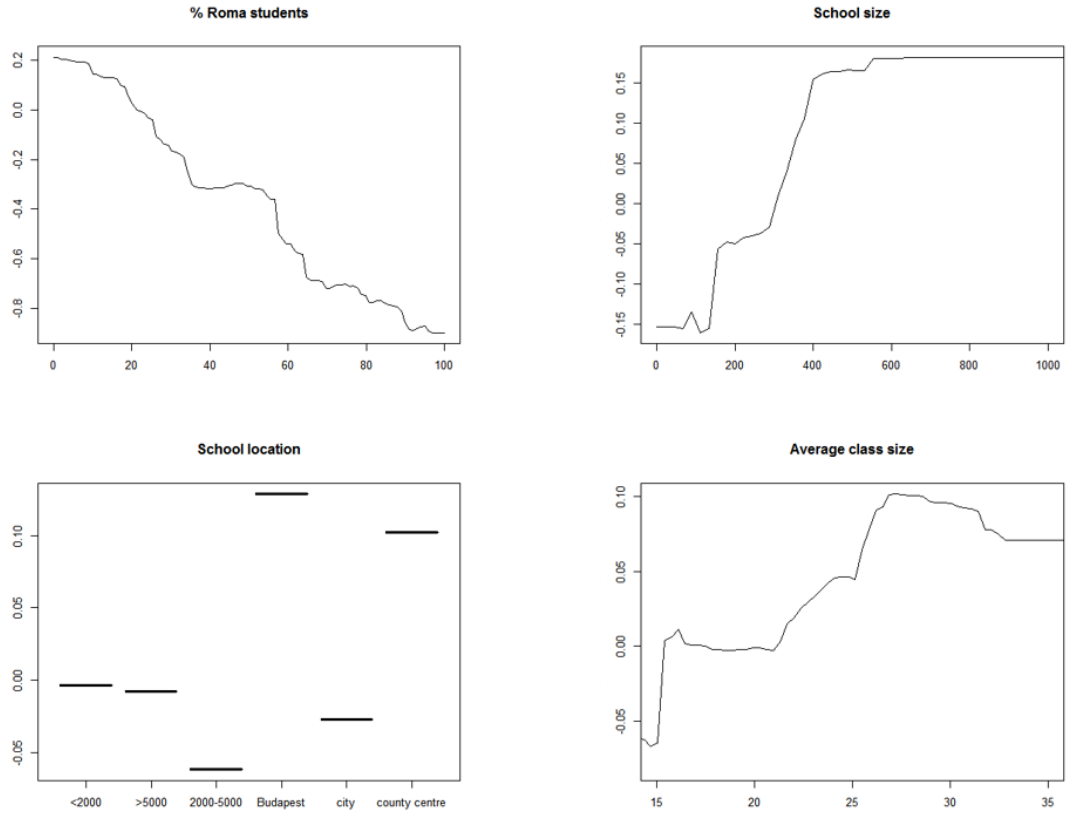
**OLS regression results.**

| Variables | Model 0 | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|---|
| School size | 0.001*** | 0.001*** | 0.001*** | 0.000 | 0.001*** | 0.001*** |
| Average class size | 0.011** | 0.011** | 0.010** | 0.0100** | 0.011** | 0.011** |
| Roma students | -0.014*** | -0.014*** | -0.012*** | -0.014*** | -0.014*** | -0.014*** |
| Number of computers | 0.003 | 0.003 | 0.002 | 0.003 | 0.003 | 0.003 |
| P age | | -0.006* | -0.006* | -0.006* | -0.006 | -0.006* |
| P experience | | 0.012*** | 0.012*** | 0.012*** | 0.013 | 0.015*** |
| P satisfaction | | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| Provider size | | 0.004 | 0.004 | 0.004 | 0.004 | 0.008** |
| **Interactions** | | | | | | |
| School size x Roma students *(Reference=Village<2k)* | | | -0.000*** | | | |
| Budapest x school size | | | | 0.001 | | |
| City x school size | | | | -0.001 | | |
| County centre x school size | | | | 0.001 | | |
| Village (2k-5k) x school size | | | | 0.000 | | |
| Village (>5k) x school size | | | | -0.001 | | |
| P experience x Age | | | | | 0.000 | |
| P experience x Provider size | | | | | | -0.000 |
| **Controls** | | | | | | |
| School location | | | | | | |
| Education provider | | | | | | |
| Region | | | | | | |
| Constant | 0.631*** | 0.844*** | 0.778*** | 0.423 | 0.837** | 0.829*** |
| Observations | 2,122 | 2,122 | 2,122 | 2,122 | 2,122 | 2,122 |
| $R^2$ | 0.213 | 0.219 | 0.221 | 0.227 | 0.219 | 0.220 |

Note: *, **, *** indicate significance at the 10%, 5% and 1% level, respectively. For conciseness, standard errors are not displayed here, although available upon request. The outcome variable is the added value of schools generated by DEA and presented in Figure 3. "Provider size" indicates the number of schools affiliated to a district, as in Table 1.

*Figure 4*

**Partial plots of selected variables**




Note: partial plots represent the marginal impact of the four most important variables of the final model (see Figure 4) on the standardized response variable (school added value): the three continuous variables "% Roma students", "School size" and "Average class size" and the categorical variable "school location".

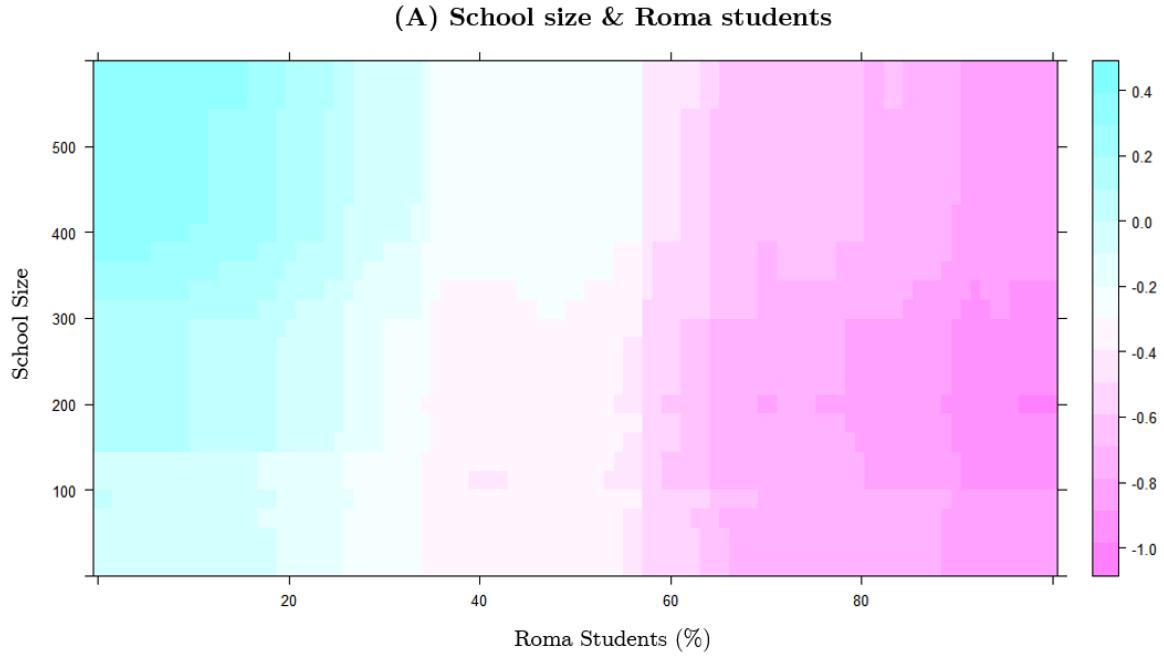
**Joint plots**. In the same way as the partial plots, the joint plots display a net effect, accounting for all other control variables in the model. In order to allow an intuitive interpretation, Boosting requires us to choose a set of two variables (components of the EPF) and display their joint plot[18]. It is important to note that the choice of this set does not affect the model outcome since *ex ante* specification of these effects is not required. In fact, the structure of the model (see Figure 1) assures that interactions within and between levels are included in the tree model anyway. As a result, no assumptions will be needed on the *existence* and *functional form* of interaction effects, as is the case when estimating a parametric education production function. In this vein, we can interpret the results illustrated in the joint plots as a data-driven estimation of complex interactions between EPF components.

[18] Selecting three variables would be possible if a 3D plot is used, despite its lower interpretability. A joint plot with more than three variables will become impossible to display.

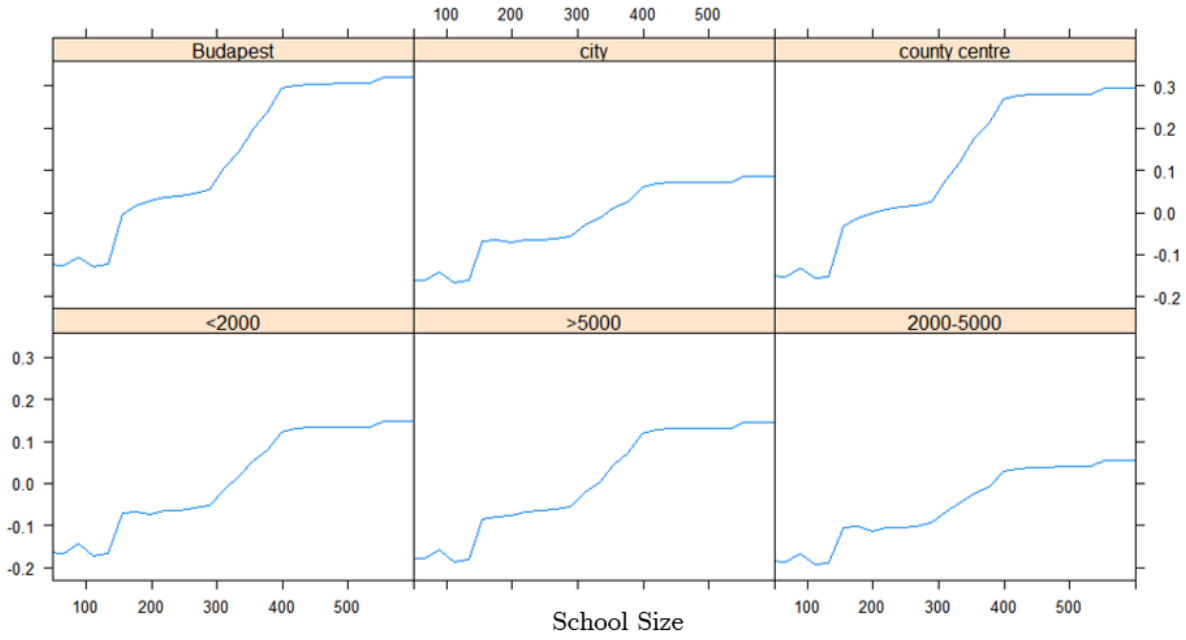
In Figure 5, we present four examples of variable combinations to illustrate the parsimonious, yet intuitive, results obtained by Boosting. Again, we allow comparison between Boosting and OLS by adding the interaction effects, displayed in Figure 5 (see Table 2, Model 2 - Model 5). Compared to Model 1, the extended models should do a better job accounting for the interactions between key variables. The results obtained when including multiplicative interactions only indicate one significant effect: *School size x Roma students*. None of the other seven interaction effects seem to be related to the added value of schools. The model improvement, in terms of $R^2$, by adding interaction effects is also very limited (less than 1%). This finding could be due to interaction effects being nonlinear. Estimating a model built on the assumption of linear interactions will then return no significant coefficients on these effects.

*Figure 5*

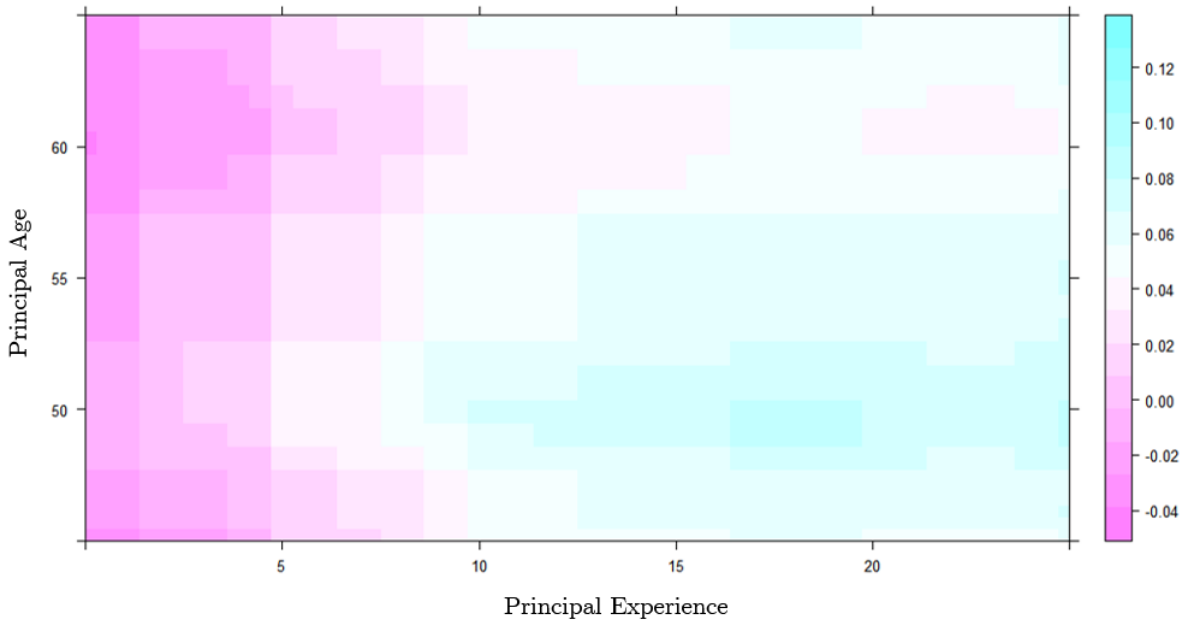
**Joint plots of selected variable combinations.**

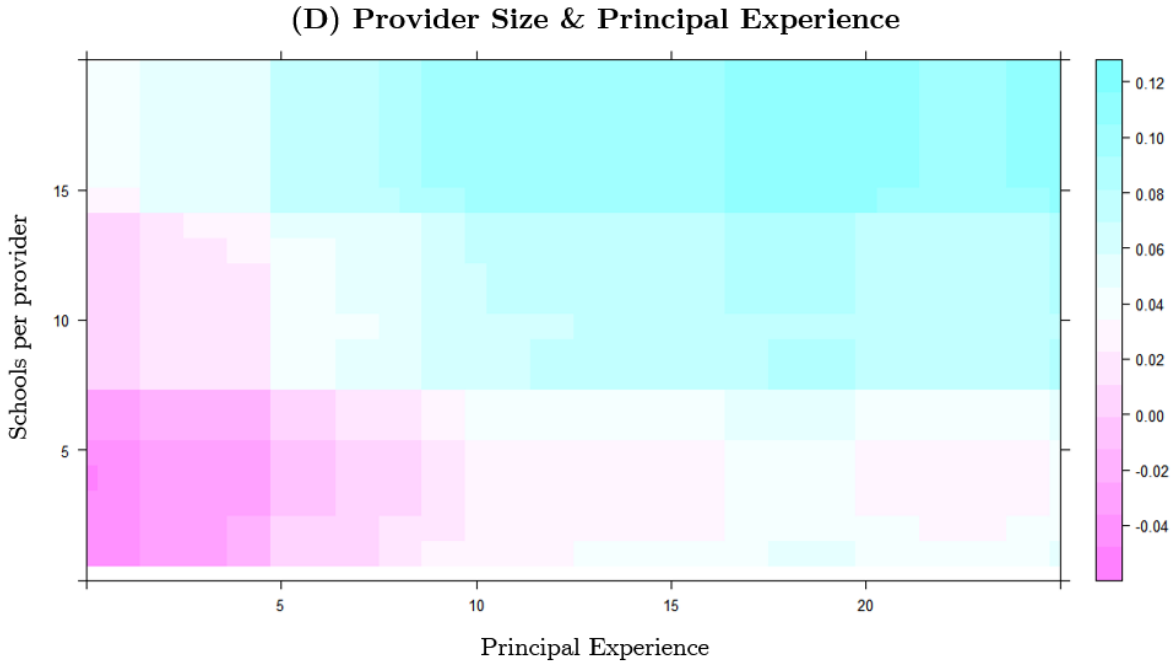


**(B) School size & school location**




**(C) Principal Age & Experience**






Note: All the above joint plots can be interpreted in the same way. The lighter the colour, the more pronounced the added value of the school at this specific point. Each point in every graph corresponds to an intersection of two values of the variables denoted on the horizontal and vertical axes. Additional joint plots and alternative variable combinations, together with the R code are available upon request.

We tackle this issue by applying the Boosting approach, outlined in Section 2, to our dataset and display the results graphically. In Figure 5, the added value of schools is indicated using a color scale and is standardized in order to ease interpretation. The vertical scales range from blue (high) to purple (low). The colors correspond to schools characterized by the intersection of two variables indicated on the axes.

In panel (A), we recognize the partial plot on the share of Roma students, as the added values of schools decreases with this share. This relationship is especially strong around the 60% level, where we observe a difference in added value of around 10% between schools with a minority of Roma students compared to 'Roma schools', again indicating the severe stigmatization in Hungarian schools. Looking at the vertical axis, schools of approximately 200 students appear to be the lowest performers (both high and low Roma schools), and increasing school size beyond 400 students does not seem to be related to higher added values, as indicated in Table 2.

If we interact both variables, we recognize the significant interaction effect *School size x Roma students* (Table 2, Model 2). We can read the gradual color change in Figure 5 in a similar way as we interpret the interaction effect: as the share of Roma students increases, the slope of school size becomes less pronounced. However, we do not impose this relationship to be linear which allows us to identify local minima and maxima. For

example, schools with the highest added value in terms of mathematics can be described as relatively large schools (400-500 students) and with very low shares of Roma students. This relationship is merely correlational and we do not claim to observe a causal effect. Nonetheless, the irregularities of interactions can reveal interesting insights into the complex education production function.

In panel (B), we interact the categorical variable 'School location' with school size. The difference in intercepts indicates differences across geographical settings while the difference in slopes indicates the importance of school size. Although the relationship in all setting looks similar, its strength can be seen on the vertical axis, indicating a strong discrepancy across locations. For example, the variation in added value related to school size differences appears to be much smaller in cities compared to county centers and Budapest. This might indicate that school size reform could have a differential impact across schools' locations, again confirming the need to account for context-dependencies when estimating the EPF. Comparing our graphical results to classical regression results reveals no significant differences across locations in Model 3, Table 2 when this relationship is assumed to be linear.[19]

The third joint plot – panel (C) – interacts principal age and experience. When looking at the vertical axis, we observe that the added value of schools is lowest for both young (40 years) and old (60 years) school principals. Schools with the highest added value appear to be led by relatively young (50 years) principals. However, when we interact the age variable with experience, we find that the added value of young principals is rather limited when they are inexperienced. Although this finding is intuitive, it reveals insights that were not captured by the linear model discussed above (Age x Experience is insignificant in Table 2, Model 4). In addition, it reveals two interesting points in panel (C). First, there seems to be some cutoff around 10 years, where the added value of schools 'jumps' to higher values. Consistent with this finding, it has been argued before that school principals "take time to realize their full effect at schools" (Coelli & Green, 2012, p.92). Possibly, this phenomenon is captured by our model in Figure 5. Also, 'optimal' schools can be identified as schools led by a 50-year old principal with around 17 years of experience in education. The difference between these schools and schools led by old (60 years) and unexperienced principals amounts to 15% SD of school added value.

To illustrate the ability of Boosting to model interactions between levels of governance, we included the interaction between provider size (number of schools per provider) and principal experience. From panel (D), we find that schools that attain the

[19] In contrast, Table C1 indicates a significantly steeper slope on school size in Budapest when reading is the outcome variable.

highest scores can be identified as schools belonging to relatively large education providers (more than 10 schools), and led by an experienced school principal. This seems to reflect the importance of a well-organized structure, which is facilitated by transparent communication (Burns & Köster, 2016), possibly captured by principal experience. Generally, the importance of experienced principals appears to be most pronounced for schools belonging to small providers. In Table 2, Model 5, the negative, yet insignificant, interaction *Principal experience x Provider size* appears to confirm this pattern. Again, we observe some cutoff values, which cannot be identified in a linear regression framework. As in panel (C), school added value 'jumps' after 10 years of experience. Also, schools belonging to providers in charge of more than 7 schools display higher value-added compared to schools in smaller districts. Larger districts might benefit from scale economies by cooperation among schools, while the added value is most pronounced if this cooperation is coordinated with sufficient skill. Interestingly, schools belonging to small providers, led by experienced principals (bottom right) do not outperform schools led by unexperienced principals, belonging to large providers (top left).

After having illustrated and interpreted the results from VI-, partial- and joint plots, we can summarize the answers to our research questions. First, we find a set of EPF components which are strongly related to the 'production' of educational outcomes in Hungarian upper secondary schools: from the socioeconomic composition of schools (as measured by % of Roma students), to some variables related to principals' decisions (school and class size), to key characteristics of school principals (age and experience). Second, our empirical approach clearly highlights the importance of modelling interactions between these components. While classical regression approaches fail to identify these interactions, our partial and joint plots clearly display substantial and nonlinear effects of interactions between key components. In a perspective of policy implications, our methodology produces results that are much more informative compared to classical approaches, as they are able to reveal more precise insights into the complexity of the education production function.

## 5. CONCLUSION

This paper illustrated how Machine Learning methods can be used to model and visualize the education production function. Flexibility in estimating the education production function reveals insights which cannot be obtained when applying a simple linear model. Models frequently used in the literature try to capture the complex, multi-faceted system of education by including multiplicative interactions. Machine Learning methods presented here allow for this complex structure without the need to make assumptions on the *existence* and *specification* of interaction effects.

Applying our model to very detailed Hungarian panel data (2008-2010), reveals two interesting findings. First, a simple OLS regression performs relatively well to model direct relationships between school added value and its determinants. Second, when the linear model is extended to include possible (linear) interactions within and between levels of education, it fails to identify the presence of these effects. In contrast, when applying Machine Learning methods, we find that school and principal characteristics matter for the added value of schools, though not unidimensional, but instead in a context of joint production. For example, cooperation between schools can only be linked to improved school performance if this process is accommodated by a school principal that possesses the right characteristics. Looking at these characteristics, we found that age and experience interact to identify the most effective principal. Illustrating our findings in joint plots introduces an intuitive graphical interpretation of interaction effects, indicating the policy relevance of this method (Brambor et al., 2006), and appealing to applied econometricians in their quest to identify heterogeneous effects (Mullainathan & Spiess, 2017). Despite the complexity of the model, results can be easily read, while at the same time, flexible interactions provide a more realistic insight into the (education) production function.

Further research will be needed to fine-tune our results. Other measures of school added-value can be introduced, offering an alternative to the DEA approach suggested in this paper. Also, we do not claim to provide causal evidence on the determinants of school value-added. However, despite the limited causal nature of our findings, we illustrated the benefits of ML methods over common estimation methods when modelling a complex education production function. In the same vein, Machine Learning methods will prove useful when exploiting discontinuities resulting from policy shocks to obtain causal inference (Athey & Imbens, 2017). The release of post-reform Hungarian data offers leeway for this kind of analyses.

**REFERENCES**


Athey, S., & Imbens, G. (2017). The State of Applied Econometrics - Causality and Policy Evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Bloom, N., Lemos, R., Sadun, R., & Van Reenen, J. (2015). Does Management Matter in schools? *The Economic Journal*, *125*(584), 647–674.

Blumenstock, J. E. (2016). Fighting poverty with data. *Science*, *353*(6301), 753–754.

Brambor, T., Clark, W. R., & Golder, M. (2006). Understanding interaction models: Improving empirical analyses. *Political Analysis*, *14*(1), 63–82.

Burgess, S. (2016). Human Capital and Education: The State of the Art in the Economics of Education. *IZA Discussion Paper Series*, *No. 9885*.

Burns, T., & Köster, F. (2016). *Governing Education in a Complex World*. OECD Publishing, Paris.

Chalfin, A., Danielli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and Selection of Human Capital with Machine Learning. *American Economic Review*, *106*(5), 124–27.

Coelli, M., & Green, D. A. (2012). Leadership effects: School principals and student outcomes. *Economics of Education Review*, *31*(1), 92–109.

Cortez, P., & Silva, A. (2008). *Using Data Mining To Predict Secondary School Student Performance. University of Minho.*

Gerritsen, S., Plug, E., & Webbink, D. (2017). Teacher Quality and Student Achievement: Evidence from a Sample of Dutch Twins. *Journal of Applied Econometrics*, *32*, 643–660.

Haveman, R., & Wolfe, B. (1995). The Determinants of Children's Attainments: Findings and Review of Methods. *Journal of Economic Literature*, *33*(4), 1829–1878.

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

Kertesi, B. G., & Kezdi, G. (2011). The Roma / Non-Roma Test Score Gap in Hungary. *American Economic Review (Papers & Proceedings)*, *101*(3), 519–525.

Loveless, T., & Hess, F. M. (2007). Introduction: What Do We Know about School Size and Class Size. *Brookings Papers on Education Policy*, *2006*(1), 1–14.

Ma, X. (2005). Growth in mathematics achievement: analysis with classification and regression trees. *The Journal of Educational Research*, *99*(2), 78–86.

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

OECD. (2010). *OECD Review on Evaluation and Assessment Frameworks for*

*Improving School Outcomes - Hungary Country Background Report*. Paris.

Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, *136*(1), 31–64.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining students-opinion data with regression and decision tree analysis. *Research in Higher Education*, *45*(3), 251–269.

Todd, P. E., & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal*, *113*(485), F3-33.

Vanthienen, J., & De Witte, K. (2017). *Data Analytics Applications in Education*. Taylor & Francis.

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

## ONLINE APPENDIX A: ESTIMATING THE ADDED VALUE OF SCHOOLS USING DEA

Assume there are data on *N* inputs and *M* outputs for each of *I* units. For the i-th unit these are represented by the column vectors $x_i$ and $q_i$ respectively. The *N*x*I* input matrix, *X*, and the *M*x*l* output matrix, *Q*, represent the data for all *I* units. An intuitive way to introduce DEA is via the *ratio* form. For each unit, we would like to obtain a measure of the ratio of all outputs over all inputs, such as $\frac{u'q}{v'x}$, where $u$ is an *M*x*l* vector of output weights and $v$ is a *N*x*I* vector of input weights. The optimal weights are obtained by solving the mathematical programming problem:

$$\max_{u,v}\left(\frac{u'q_i}{v'x_i}\right),$$
$$\text{s.t.}$$
$$\frac{u'q_i}{v'x_i} \leq 1 \text{ with } i = 1, \dots, I$$
$$u, v \geq 0.$$

This involves finding values for $u$ and $v$, such that the efficiency measure for the *I*-th unit is maximized, subject to the constraints that all efficiency measures must be less than or equal to 1. One problem with this particular ratio formulation is that it has an infinite number of solutions. To avoid this, one can impose the constraint $v'x_i = 1$, which can be written as:

$$\max_{\mu,\eta}(\mu' q_i),$$
$$\text{s.t.}$$
$$\eta' x_i = 1\,,$$
$$\mu' q_i - \eta' x_i \leq 1 \text{ with } i = 1, \dots, I$$
$$\mu, \eta \geq 0.$$

where the change of notation from $u$ and $v$ to $\mu$ and $\eta$ is used to stress that this is a different linear programming problem.

To compute the added value of schools, we apply DEA with output-orientation, using the following set of variables:[20]

- Inputs: school average test score (math or reading) at grade 6, school average socio-economic status index;
- Outputs: school average test score (math or reading) at grade 8.

[20] Note that the NABC dataset has been used at the school level in this study. As a result, all variables are defined at the school level, both in the first and second stage.

In addition, we release the assumption of constant returns to scale by adding the restriction $\sum \mu = 1$ on the weights. This allows a more realistic representation of school added value, where the school average test score in year 6 does not necessarily move in a linear way with the school score in year 8. It might for example be the case that schools 'starting' at a higher average score attract better teachers, resulting in a faster learning process. If this were true, imposing linearity on the 'production function' will bias our estimates of school added value.

We run two separate models, one looking at the mathematics test scores and the other at the reading scores. Although one key advantage of DEA is to handle multiple inputs and outputs simultaneously, we want to check whether the structure of the EPFs (i.e. the factors associated with output production) are different across subjects.

## ONLINE APPENDIX B: THE HUNGARIAN EDUCATION SYSTEM

The dataset we use for explaining the potential of the proposed methodology in an empirical setting deals with upper secondary schools in Hungary.[21] Following a trend towards decentralization until 2013, the Hungarian educational system was characterized by a high degree of local autonomy. The law of 2011/CXC. on Public Education, effective from September 2013, has centralized the system, and introduced several changes. Before the 2013 reform, more powers were attributed to school management (principal and provider), emphasizing their responsibility over results. Our analysis will refer to the pre-2013 decentralized system due to this fact and the lack of more current data. The hierarchical structure is presented in Figure A2.

In this paper, schools are the unit of our analysis. Schools can consist of different 'sites' which belong to the same school but are separate administrative units.22 Most schools coincide with their site (around 80%). Schools are governed by a principal (and deputy principals if there are multiple sites), who manages a team of teachers, and are supervised by an education provider. The provider of education (usually the local government23) coordinates the set of schools under its supervision.

[21] This study is part of the Horizon 2020 Twinning project 'Economics of Education Network (EdEN)', an enhanced cooperation in the field of education economics between three economics of education research groups in EU-15 countries – KU Leuven (Belgium), U Maastricht (Netherlands) and Politecnico di Milano (Italy) – and the CERSHAS (Hungary).
[22] For example, a school can choose to divide its activities into two separate sites, one providing primary schooling and the other providing secondary schooling.
[23] The local government in Hungary is similar to the English Local Education Authority, but deals with several other public issues besides education.

*Figure A 1*

**Schematic representation of the Hungarian education system**

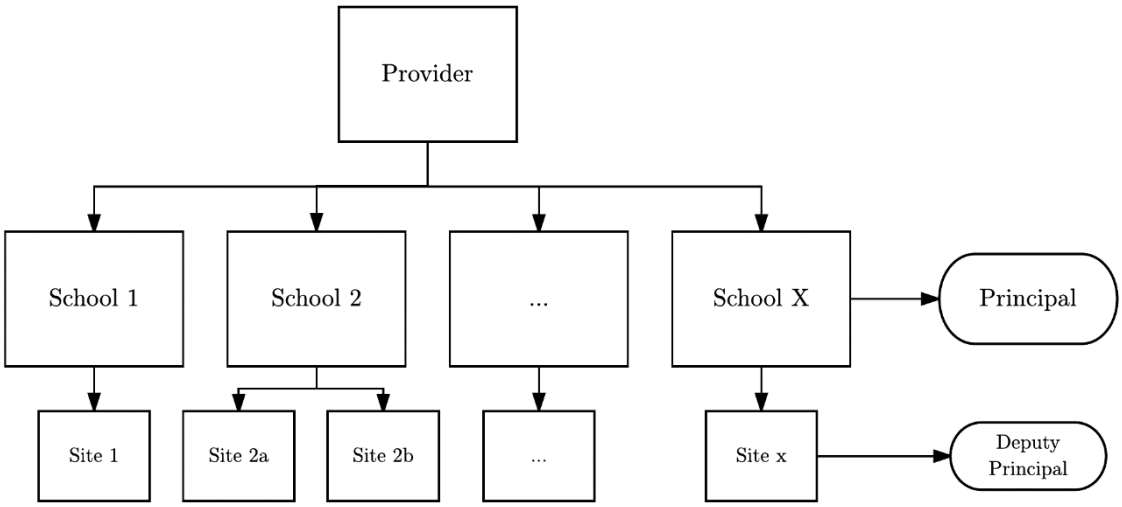
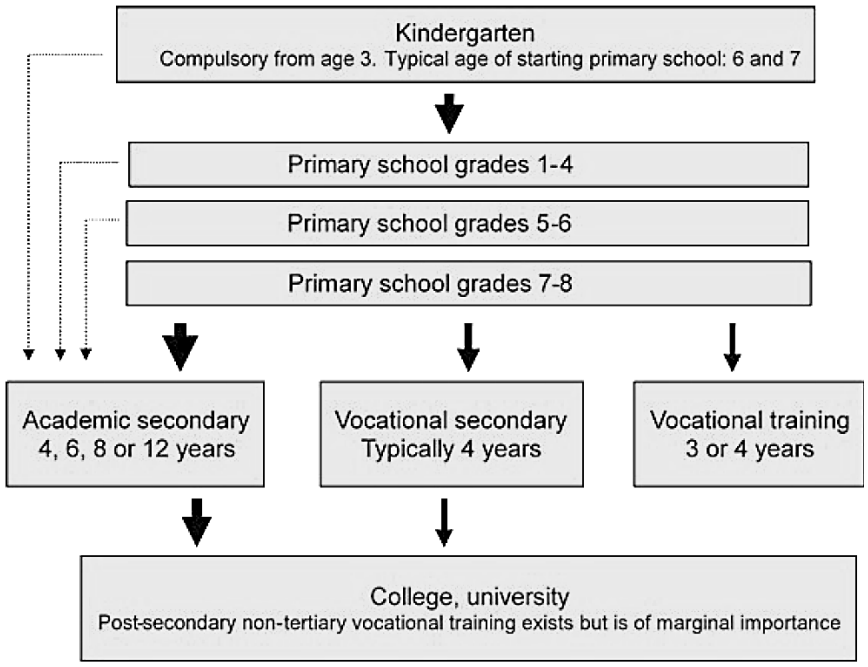



*Figure A 2*

**The structure of education in Hungary**




While most of the education funding in Hungary, before 2013, was covered by a per-student lump-sum grant provided by the central government to each education provider (for private as well as the public providers), these grants were insufficient to cover all

costs. Hence, the providers of education had to subsidize their schools according to their needs or its financial capacities. The providers of education were also the official employers of teachers and school principals before 2013. In other words, they were crucial players in every aspect of the schools' life. All in all, the pre-2013 Hungarian education system was one of the most decentralized within the OECD (see OECD, 2010), which means school providers, and principals could use this autonomy to tailor their school organization policies to the specific needs of their students.

Table A 1 (Appendix) shows which students are measured within NABC. There are several explicit goals of this assessment. First, the major goal is to provide more detailed and more frequent feedback for the educational policy relative to international surveys. The second is to offer a tool for education providers and schools to improve their performance. The third goal is to set the grounds for a future accountability system and provide higher transparency. In addition to all this, it offers invaluable data for researchers to address education puzzles. Unfortunately, up until 2008 the database could only be analysed on a cross sectional basis, because it did not contain permanent student level identification numbers. From 2008 onwards the biannual datasets are linked by student IDs, allowing more detailed (value-added) analyses.

In addition to mathematics and literacy test scores, which are common to education datasets, our database contains extensive information on the student background, principals, and the school. This paper uses data on the 2008/6$^{th}$ grade cohort. All of the students were observed two years later, in 8$^{th}$ grade (2010). We purposely selected these two moments in time to minimize possible biased results resulting from students changing schools. As can be seen in Figure A2, most students go to 8 years of primary and lower secondary education (*általános iskola*). Less than 10% of each cohort exit this 8 years of comprehensive training to enter an elite, academic track. The exit points are at age 10 (after grade 4) and at age 12 (after grade 6). As there are no exit points between grade 6 and 8 student mobility between schools is very low between these two points in time. Hence, focusing on the 6$^{th}$ and 8$^{th}$ grade student achievement allows us to disentangle the added value of schools.

*Table A 1*

## The official NABC database

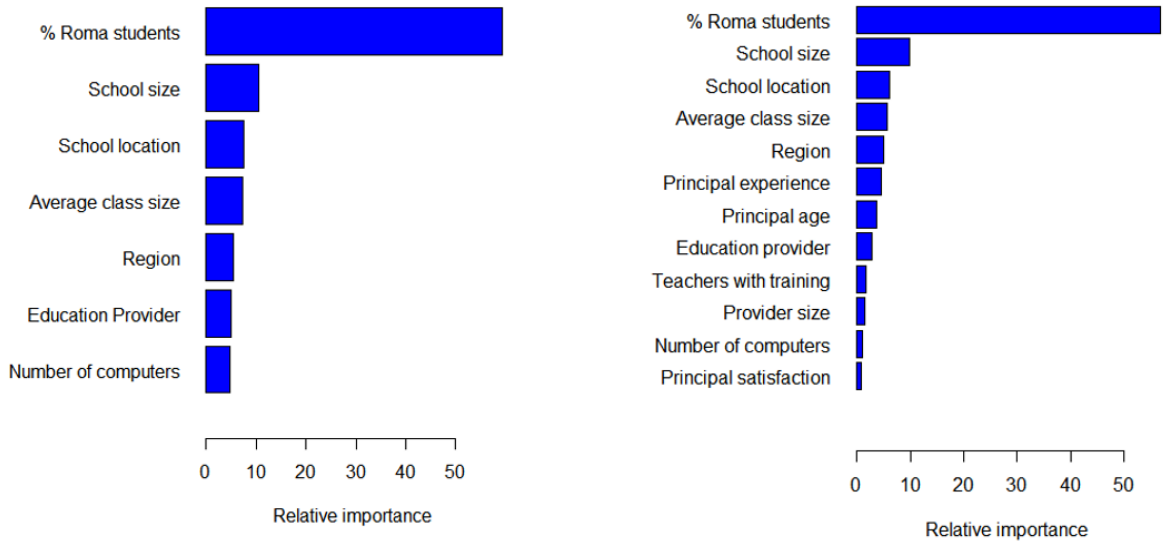
| | 6th grade | 8th grade | 10th grade |
|---|---|---|---|
| **2003** | 20 students from every school | 0 | 20 students from each track from each school |
| **2004** | 20 students from every school | 20 students from every school | 20 students from each track from each school |
| **2006** | every student from a sample of 195 schools | full cohort | 30 students from each track from each teaching site |
| **2007** | every student from a sample of 200 schools | full cohort | 30 students from each track from each teaching site |
| **2008*** | full cohort | full cohort | full cohort |
| **2009*** | full cohort | full cohort | full cohort |
| **2010*** | full cohort | full cohort | full cohort |
| **2011*** | full cohort | full cohort | full cohort |
| **2012*** | full cohort | full cohort | full cohort |
| **2013*** | full cohort | full cohort | full cohort |
| **2014*** | full cohort | full cohort | full cohort |

* Permanent individual identification numbers are available

## ONLINE APPENDIX C: READING RESULTS

*Figure C1*

**VI plot of base model (left) and final model (right) for reading**




Note: base and final models differ in the predictors involved. Base model uses 7 predictors (the ones appearing on the left panel of the figure), while the final model uses 12 predictors (the ones appearing on the right panel of the figure). It is worth to notice that Random Forest identifies the same important variables.

*Table C1*

**OLS regression results for reading**

| Variables | Model 0 | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|---|
| School size | 0.001*** | 0.001*** | 0.001*** | -0.000 | 0.001*** | 0.001*** |
| Average class size | 0.015*** | 0.015*** | 0.015*** | 0.014*** | 0.015*** | 0.015*** |
| Roma students | -0.020*** | -0.020*** | -0.017*** | -0.020*** | -0.020*** | -0.020*** |
| Number of computers | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| P age | | -0.004 | -0.004 | -0.004 | -0.001 | -0.004 |
| P experience | | 0.012*** | 0.011*** | 0.012*** | 0.039 | 0.012*** |
| P satisfaction | | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| Provider size | | 0.004 | 0.004 | 0.004 | 0.004 | 0.005 |
| **Interactions** | | | | | | |
| *(Reference=Village<2k)* | | | -0.000*** | | | |
| Budapest x school size | | | | 0.002** | | |
| City x school size | | | | 0.000 | | |
| County centre x school size | | | | 0.001* | | |
| Village (2k-5k) x school size | | | | 0.001 | | |
| Village (>5k) x school size | | | | 0.000 | | |
| P experience x Age | | | | | -0.000 | |
| P experience x Provider size | | | | | | -0.000 |
| **Controls** | | | | | | |
| School location | | | | | | |
| Education provider | | | | | | |
| Region | | | | | | |
| Constant | 0.495** | 0.594** | 0.499* | 0.234 | 0.423 | 0.590** |
| Observations | 2,122 | 2,122 | 2,122 | 2,122 | 2,122 | 2,122 |
| $R^2$ | 0.327 | 0.332 | 0.336 | 0.339 | 0.333 | 0.332 |

Note: *, **, *** indicate significance at the 10%, 5% and 1% level, respectively. For conciseness, standard errors are not displayed here, although available upon request. The standardized outcome variable is the added value of schools generated by DEA and presented in Figure 3. “Provider size” indicates the number of schools affiliated to a district, as in Table 1. ‘P’ is short for Principal.