The Value Added of Machine Learning to Causal Inference: Evidence from Revisited Studies*

Anna Baiardi† and Andrea A. Naghi‡

Abstract: A new and rapidly growing econometric literature is making advances in the problem of using machine learning methods for causal inference questions. Yet, the empirical economics literature has not started to fully exploit the strengths of these modern methods. We revisit influential empirical studies with causal machine learning methods aiming to connect the econometric theory on these methods with empirical economics. We focus on the double machine learning, causal forest and generic machine learning methods, in the context of both average and heterogeneous treatment effects. We illustrate the implementation of these methods in a variety of settings and highlight the relevance and value added relative to traditional methods used in the original studies.

Keywords: machine learning, causal inference, average treatment effects, heterogeneous treatment effects.

J.E.L. Classification: C01, C21, D04

---

*Baiardi acknowledges support from EU Horizon 2020, Marie Skłodowska-Curie individual grant (No. 840319). Naghi acknowledges support from EU Horizon 2020, Marie Skłodowska-Curie individual grant (No. 797286). Financial support from the United Nations Sustainable Development Funds is also gratefully acknowledged. We thank participants at the Machine Learning for Economics Workshop (at Barcelona GSE Summer Forum 2019), the Netherlands Econometrics Study Group Meeting 2020, and seminar participants at University of Amsterdam and Etla Economics Research for very helpful comments. Nadja van't Hoff, Olivier Mulkin and Christian Wirths provided excellent research assistance.

†Department of Economics, Erasmus University and Tinbergen Institute. Email: baiardi@ese.eur.nl.

‡Department of Econometrics, Erasmus University and Tinbergen Institute. Email: naghi@ese.eur.nl.

1 Introduction

One of the key goals of empirical research in economics is to estimate the causal effect of a variable of interest on a targeted outcome. To avoid biases in the coefficients of interest due to omitted variables, particularly in observational studies, it is often desirable to include in the regressions a large number of controls. Even if the number of raw covariates is relatively small, including interactions and transformations can quickly increase the number of controls in the regression.

Machine learning (ML) methods can potentially be useful in such settings. However, *standard* ML prediction models are aimed for fundamentally different problems than most of the empirical work in economics. ML methods are designed and optimized for predicting the outcome in a test sample. Thus, a model is selected by optimizing the goodness of fit on the held-out test set. In contrast, in empirical economic research, the goodness of fit of a model is oftentimes reduced when estimating a causal effect, and the predictive accuracy is sacrificed in order to learn more deeply about a fundamental relationship that can guide policy decisions and counterfactual predictions (Athey and Imbens, 2019). These fundamental differences will eventually generate biased estimates if *standard* ML techniques, designed for prediction, are used in the context of causal inference.[1] Nevertheless, a new and rapidly growing econometric literature is making advances in the problem of using ML methods for causal inference questions (see, e.g., Chernozhukov et al., 2018a; Athey et al., 2018; Wager and Athey, 2018; Chernozhukov et al., 2018b). This literature brings in new insights and theoretical results that are novel for both the ML and the econometrics/statistics literature. Despite these advances, the empirical economics literature has not started yet to fully exploit the strengths of these new modern causal inference methods.

The aim of this paper is to present empirical researchers evidence regarding the merits of causal machine learning methods in realistic settings. To this end, we revisit a number of influential papers by applying causal ML methods and compare the results with the traditional methods used in the original studies. In our analysis, we focus on both the average treatment effect (ATE) and heterogeneous treatment effects (HTE). Our main contribution is to illustrate how causal

[1] One of the underlying reasons is that, for instance, high dimensional regression adjustments such as lasso, ridge, elastic net etc., shrink the estimated effects by construction, and ignoring this shrinkage will lead to biased treatment effect estimates.

ML methods can be implemented in a variety of settings, and to highlight the relevance and additional gains that causal machine learning methods bring to the table relative to the standard econometric approaches. We further support some of our main findings with several Monte Carlo simulations, where the true data generating process is known. This allows us to compare the finite sample performance of causal ML estimators with traditional estimators in settings similar to the revisited studies.

When interested in the ATE, we employ the double/debiased machine learning (DML) method of Chernozhukov et al. (2017); when the focus is on heterogeneous treatment effects (HTE), we work with the causal forest method of Athey et al. (2019) and Wager and Athey (2018), and with the generic machine learning method for heterogeneous treatment effects developed by Chernozhukov et al. (2018b). These are newly developed causal machine learning methods with well-established theoretical properties. We re-examine a set of relatively recent influential studies that span a variety of topics in applied economics, published in the following journals: *The Quarterly Journal of Economics, American Economic Journal: Macroeconomics, American Economic Journal: Applied Economics.* We choose papers for which the full replication data set is available either on the journal's website or on the authors' website. For the ATE, we revisit two observational studies: the study of Djankov et al. (2010) on the effect of corporate taxes on investment and entrepreneurship, and the paper by Nunn and Trefler (2010) on the effect of skill-biased tariffs on long-term economic growth. For the HTE, we select one observational study and one randomized control trial: we extend the observational study by DellaVigna and Kaplan (2007), which investigates the effect of Fox News on the Republican vote share, and the analysis by Loyalka et al. (2019) on the effect of a teacher training randomized intervention on student performance. All these papers include careful econometric analyses of the main research question and mechanisms, which we do not aim to re-examine in full. We instead focus on analyzing the main questions.

Based on our results from the sample of revisited paper, we derive and systematize four main reasons why causal machine learning methods are relevant for causal analysis and add value relative to the traditional methods. These are general reasons that are not only valid for the specific settings or datasets of the papers that we revisit.

Firstly, causal ML methods are powerful tools in using data to recover complex interactions among variables and flexibly estimate the relationship between the outcome, the treatment and the covariates. This feature is key when drawing inference based on the assumption that the treatment is unconfounded conditional on the observables, as in the case of most of the revisited studies, since this assumption is not testable. As some covariates can be correlated with both the treatment variable and the outcome, failing to condition on all relevant confounders may lead to biased estimates for the treatment effect. For example, for the effect of corporate taxes on investment and entrepreneurship, the original analysis in Djankov et al. (2010) shows a negative and significant effect of corporate taxes on investment and entrepreneurship, but the authors show that these results do not survive when conditioning on all the potential controls at once. However, when implementing DML, we obtain larger estimates compared to Djankov et al. (2010), which are often statistically significant. Furthermore, our analysis of the effect of skill-biased tariffs on growth suggests a smaller effect compared to Nunn and Trefler (2010), which is often not statistically significant. We thus argue that the DML estimates are more robust to potential nonlinear confounders.[2]

Secondly, causal ML methods can be more suited than traditional methods when the number of covariates is large relative to the sample size, as they assume that the model is sparse, (i.e., only a small number of covariates are relevant), and they use regularized regressions. For instance, in the study by Djankov et al. (2010) and in some of the specifications in Nunn and Trefler (2010), the number of raw covariates is large compared to the sample size, thus taking into account all possible nonlinear terms, such as interactions and transformations, would not be possible when using traditional methods. Indeed, no nonlinear terms other than logarithms are considered in Nunn and Trefler (2010), and no nonlinear terms are included in Djankov et al. (2010). In contrast, by using the DML method we ensure that our results take into account all potentially relevant confounders at once, both linearly and nonlinearly.

Thirdly, the use of causal ML methods allows for *systematic model selection.*

[2]It is important to note here that the idea of estimating treatment effects without making parametric assumptions about the way in which the covariates enter the equation has already been considered in the semiparametric econometrics literature (see the review paper of Imbens and Wooldridge, 2009, and Imbens and Rubin, 2015). However, in practice, these semiparametric kernel methods quickly break down if they have to deal with more than a few covariates.

Many ML methods search for the best functional forms by estimating and comparing a wide range of alternative model specifications; the model selection is thus data-driven and fully documented. For example, our results for the effect of corporate taxes, originally explored by Djankov et al. (2010), show that the data-driven model selection implemented by DML, which keeps a smaller set of influential confounding factors from among a large set of potential controls, leads to larger coefficients in absolute value and lower standard errors compared to OLS regressions where all the covariates are included. With the traditional approach to model selection, uncertainty about the correct specification of the model can lead to choices that are relatively *ad hoc*; different specifications may lead to different point estimates, which in turn may lead to different policy decisions. Moreover, we further illustrate how these methods are also very useful tools for *supplementary analyses* or *robustness checks*. Typically, supplementary analysis is performed by presenting a number of selected regression specifications, while the approach of causal ML methods is more systematic, and ensures that important transformations of covariates that are not considered relevant a priori are not missed. For instance, we can consider our analysis of Nunn and Trefler (2010) as a robustness check, as with DML we control for a data-driven function of the covariates. In this case, our results are different from the original analysis and statistical significance is lost.

Finally, causal machine learning methods prove to be very useful when one is interested in estimating heterogeneous treatment effects. As causal ML methods can handle many covariates potentially responsible for treatment effect heterogeneity in a systematic way, it is less likely that relevant heterogeneous effects will be missed, compared to manually modelling different interaction terms. This feature is exemplified by our analysis of the heterogeneous effects of Fox News on the Republican vote share first explored by DellaVigna and Kaplan (2007) and of the teacher training intervention studied by Loyalka et al. (2019): our results reveal drivers of heterogeneity that were unexplored in the original analysis. In addition, causal ML methods tailored for estimating heterogeneous treatment effects provide valid confidence intervals in high dimensional settings, as opposed to traditional methods where standard p-values for single hypothesis testing are not reliable.[3] This is due to the multiple hypothesis testing problem, which can occur when researchers search iteratively for treatment effect heterogeneity, over a large number of covariates.[4,5]

Our main findings from the revisited studies are supported by several Monte Carlo simulations inspired by empirically relevant settings. Our focus is on evaluating the relative finite sample performance of traditional and causal machine learning methods: i) when the relationship between the outcome and the covariates, as well as the treatment and the covariates, is either linear or nonlinear and; ii) when the number of covariates used in estimation increases relative to the sample size. To this end we focus on the DML and show that it outperforms OLS when the true nuisance relationship is nonlinear. Moreover, we find that the performance of DML relative to OLS improves as the number of covariates increases relative to the sample size, in both the linear and the nonlinear case.

The econometric theory literature on adapting standard machine learning techniques to causal inference questions is by now fast growing. See for example Chernozhukov et al. (2017), Chernozhukov et al. (2018a), Athey et al. (2018), Farrell et al. (2021), Colangelo and Lee (2020) for the ATE; and Athey and Imbens (2016), Wager and Athey (2018), Athey et al. (2019), Chernozhukov et al. (2018b), Semenova et al. (2018), Oprescu et al. (2019) for the HTE. In the statistics literature, estimation of ATE and HTE with machine learning methods has been the focus in Hill (2011), Imai et al. (2013), Van der Laan and Rose (2011), Su et al. (2009), Zeileis et al. (2008), among others. A few papers started employing the above mentioned methods in interesting early applications. See for example, Davis and Heller (2020), Davis and Heller (2017), Knaus et al. (2020), Strittmatter (2019) and Bertrand et al. (2017) for the causal forest, and Deryugina et al. (2019) for

[3]Note that the causal forest method by Wager and Athey (2018) is not developed for very high dimensional settings; however, the generic machine learning method of Chernozhukov et al. (2018b) can handle a large number of covariates.

[4]While solutions have been proposed to correct for the issue of multiple hypothesis testing (for example, List et al., 2016), when the number of covariates is large, the power of these approaches to detect heterogeneity is low (Athey and Imbens, 2017).

[5]A related issue is the ex-post selection of significant heterogeneous effects. To avoid this problem, in randomized control trials researchers are often required to specify before the experiment which heterogeneous effects they are interested to look into, in order to avoid searching for, and only reporting, significant effects. However, this limits the ability of the researcher to find unexpected relevant heterogeneity. Causal ML methods ensure that relevant heterogeneity is not missed while also providing valid confidence intervals. In addition, in observational studies, where pre-analysis plans are not common practice, causal ML methods can be particularly useful.

the generic machine learning.

In what follows, we present our main findings on average treatment effects using double machine learning in Section 2. The analysis of heterogeneous treatment effects using the causal forest and the generic machine learning method are described in Section 3. In Section 4 we summarize our main takeaways and recommendations for the applied researchers interested in employing these methods. An intuitive description of the methodology, more details on the revisited papers, details on the implementations of the methods, and the results of the Monte Carlo study are deferred to the Online Appendix.

2 Average Treatment Effects

This section contains the analysis on the ATE for the effect of corporate taxes on investment and entrepreneurship (Djankov et al., 2010) and the effect of skill-biased tariffs on growth (Nunn and Trefler, 2010) using the double machine learning method (Chernozhukov et al., 2017).

2.1 The Effect of Corporate Taxes on Investment and Entrepreneurship

Description of Original Analysis. The first paper that we revisit using causal machine learning methods investigates the relationship between corporate taxes on investment and entrepreneurship (Djankov et al., 2010). This is an observational study that shows a negative effect of corporate taxes on investment and entrepreneurship, by estimating OLS country-level regressions with different measures of corporate tax rates for the year 2004. The sample includes a set of 50-85 countries, depending on the specification. In the original paper, four outcome variables are examined: investment as a percentage of GDP, FDI as a percentage of GDP, business density per 100 people, and the average entry rate. Three measures of corporate taxes are considered: statutory corporate tax rates, actual first-year corporate income tax liability of a new company, and the tax rate which takes into account actual depreciation schedules going five years forward.

The original paper reports the results for several regression specifications with different sets of control variables, to account for potential confounders that correlate with corporate tax rates, and are also determinants of the outcomes.[6] Djankov

[6]The first set of controls includes measures of other taxes; the second set includes measures

et al. (2010) present regression results where the first three sets of covariates are added separately. A final robustness check includes all control variables (12 in total) in the same regression. In the specifications which include only one set of controls at a time, the paper shows a negative and statistically significant effect of corporate taxes on entrepreneurship and investment. However, when adding all the controls, the relationship is still negative, but the coefficients are smaller in size and no longer statistically significant.

DML Analysis. We revisit the final robustness check of the paper, which includes all four sets of covariates at the same time, using the DML partially linear model. Table 1 presents the results. Columns (1) to (7) display the DML point estimates for the effect of corporate taxes on investment and entrepreneurship, using different ML methods to estimate the nuisance functions. Further details on how the DML estimates are obtained, the methods used and the tuning parameters are described in Section B.1 of the Online Appendix.

We notice that all the DML point estimates have negative signs and generally similar magnitudes across the ML methods. Compared to the original paper results with the full set of covariates, reported in column (8), the magnitude of the DML coefficients is higher in absolute value, and the standard errors are lower in most regressions. Additionally, the results are statistically significant, at least at the 10% level, in half (42 out of 84) of the regressions.

It seems that applying regularization here leads to lower standard errors and higher precision. However, in the absence of the known ground truth, whether the DML estimates are closer to the truth or not can be questioned. To offer further clarifications on this point, we note that both the original analysis and our analysis rely on the unconfoundedness assumption. In the case of the OLS analysis, the (implicit) assumption is that it is sufficient to control for all factors linearly. However, DML allows for a more flexible estimation, including potential nonlinear confounders as well as linear controls. This means that DML allows to relax the original assumption and replaces it with a weaker assumption, i.e., that the effect of confounders can be sufficiently controlled for by including the same controls as in the original analysis both linearly and nonlinearly. Furthermore, one

---

for the number of other tax payments made and for tax evasion; the third set includes measures for institutions; the fourth set includes measures of inflation. Section B.1 of the Online Appendix includes more details on the regressions estimated in Djankov et al. (2010) and describes the control variables.

Table 1: The Effect of Corporate Taxes on Investment and Entrepreneurship

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) OLS
	Panel A: Investment 2003-2005							
Statutory corporate tax rate	-0.074	-0.069	-0.068	-0.07	-0.056	-0.066	-0.071	-0.064
	(0.09)	(0.072)	(0.076)	(0.087)	(0.102)	(0.087)	(0.088)	(0.098)
First-year effective tax rate	-0.114	-0.129	-0.154	-0.144	-0.122	-0.13	-0.133	-0.117
	(0.094)	(0.087)	(0.093)	(0.096)	(0.097)	(0.092)	(0.095)	(0.106)
Five-year effective tax rate	-0.187	-0.182	-0.211	-0.21	-0.217	-0.216	-0.207	-0.189
	(0.089)	(0.089)	(0.092)	(0.097)	(0.103)	(0.095)	(0.101)	(0.118)
Observations	61	61	61	61	61	61	61	61
	Panel B: FDI 2003-2005							
Statutory corporate tax rate	-0.148	-0.157	-0.153	-0.14	-0.085	-0.133	-0.114	-0.030
	(0.083)	(0.086)	(0.092)	(0.094)	(0.093)	(0.088)	(0.09)	(0.066)
First-year effective tax rate	-0.141	-0.194	-0.178	-0.157	-0.136	-0.161	-0.137	-0.1
	(0.091)	(0.081)	(0.081)	(0.074)	(0.078)	(0.08)	(0.079)	(0.071)
Five-year effective tax rate	-0.147	-0.177	-0.167	-0.165	-0.139	-0.157	-0.14	-0.095
	(0.084)	(0.073)	(0.074)	(0.077)	(0.082)	(0.077)	(0.076)	(0.081)
Observations	61	61	61	61	61	61	61	61
	Panel C: Business density							
Statutory corporate tax rate	-0.062	-0.092	-0.069	-0.07	-0.056	-0.066	-0.06	-0.034
	(0.066)	(0.072)	(0.061)	(0.063)	(0.077)	(0.069)	(0.064)	(0.083)
First-year effective tax rate	-0.104	-0.156	-0.124	-0.122	-0.105	-0.114	-0.1	-0.068
	(0.076)	(0.082)	(0.07)	(0.069)	(0.085)	(0.072)	(0.07)	(0.092)
Five-year effective tax rate	-0.091	-0.139	-0.122	-0.107	-0.115	-0.114	-0.104	-0.070
	(0.076)	(0.08)	(0.071)	(0.067)	(0.087)	(0.074)	(0.075)	(0.103)
Observations	60	60	60	60	60	60	60	60
	Panel D: Average entry rate 2000-2004							
Statutory corporate tax rate	-0.112	-0.147	-0.141	-0.127	-0.067	-0.112	-0.106	-0.029
	(0.073)	(0.068)	(0.064)	(0.065)	(0.084)	(0.067)	(0.069)	(0.086)
First-year effective tax rate	-0.130	-0.144	-0.143	-0.125	-0.131	-0.126	-0.117	-0.083
	(0.072)	(0.064)	(0.065)	(0.066)	(0.086)	(0.07)	(0.072)	(0.094)
Five-year effective tax rate	-0.154	-0.153	-0.164	-0.164	-0.191	-0.168	-0.167	-0.133
	(0.084)	(0.069)	(0.07)	(0.07)	(0.091)	(0.08)	(0.077)	(0.103)
Observations	50	50	50	50	50	50	50	50
Raw covariates	12	12	12	12	12	12	12	12

Notes: Analysis of Table 5D of Djankov et al. (2010) using DML. Column 8 reports the original paper estimates. Standard errors are reported in parentheses. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. The number of covariates does not include the treatment variable.

might be interested in investigating what are these nonlinear terms that make the estimates different. However, this can be a challenging task when ML methods (such as neural networks, hybrid methods etc.) are used to estimate the nuisance functions. What can potentially be done is analyzing the lasso coefficients that are not shrunk to zero and looking for nonlinearities among these. As an example, we show in Figure C.1, in the Online Appendix, the most relevant among the nonlinear terms selected by the lasso, for one of the DML regressions reported in Table 1.[7]

[7] It is important to note here that we do not make inference using the lasso coefficients, but

Here, we note that several nonlinear terms appear in both the treatment nuisance function $\hat{m}(\cdot)$ and in the outcome nuisance function $\hat{g}(\cdot)$.[8] This is suggestive of the fact that there are nonlinearities that are correlated with both the treatment variable and the outcome. These were missed by the analysis in the original paper, and their omission could lead to biased coefficients of the corporate taxes variables. In this case, controlling for all relevant confounders strengthens the main results of the original analysis: in many cases the DML treatment effect estimates are larger in absolute value, and statistically significant. The results of our Monte Carlo simulations, presented in Section D of the Online Appendix, further highlight the relevance of using DML in the presence of nonlinearities, even is small sample sizes.

The DML results are obtained by tuning the parameters of the ML methods via cross-validation, whenever this is theoretically justifiable. Some of the parameters, however, are not data-driven (for example the number of trees, or the leaf node size). Thus, we perform additional sensitivity checks on the values used for these non-adaptive tuning parameters. In addition, we change the activation function and vary the number of layers in the neural net. The results, not reported, but available upon request, are consistent with those reported here.

The good performance of causal machine learning methods is subject to the assumption of *sparsity*. However, the sparsity assumption is not testable and thus it must be used with caution. In our empirical applications, it is reassuring however that the results obtained from the different ML methods give very close second-stage DML estimates (the estimates of the ATEs). This is consistent with the existence of a sparse basis which is concomitantly captured by all ML methods.

This empirical application is a good example to illustrate the usefulness of causal ML methods in the *typical trade-off* that applied researchers often face. On the one hand, the researcher wants to control for as many potential confounders as possible, in order to improve the credibility of the unconfoundedness assumption. On the other hand, naively controlling for a large set of covariates, especially when the sample size is small, can lead to imprecise estimates and larger standard errors. Notice that in this example, the authors implement a "kitchen sink" regression and

---

we analyze the magnitude of the coefficients as a measure of the covariates' importance for predicting the outcome and the treatment variables.

[8]Further details about the lasso coefficients analysis are reported in Section B.1 of the Online Appendix.

control for all the covariates at once, resulting in larger standard errors than the ones that we obtain. The DML method helps with this trade-off by improving the credibility of the unconfoundedness assumption (as it captures more flexibly the effect of confounders), but, at the same time, it implements a data-driven variable selection technique to keep a smaller set of influential confounding factors from among a large set of potential controls, thus resulting in lower standard errors.

Lastly, to further support our obtained causal ML results, we also perform the analysis and compute the ATEs using the causal forest.[9] We report the causal forest estimates in Table C.2 in the Online Appendix. The results are consistent with the DML estimates.[10]

2.2 The Effect of Skill-Biased Tariffs on Growth

Description of Original Analysis. The study by Nunn and Trefler (2010) investigates the relationship between skill-biased tariffs, i.e., a tariff structure that disproportionately favours skill-intensive industries, and long-term economic growth. The authors develop a theoretical framework based on Grossman and Helpman (1991) that shows how tariffs that focus on skill-intensive industries can lead to a disproportionate expansion of skill-intensive industries, which then leads to higher long-term growth. Furthermore, using both cross-country and industry level data, the paper provides evidence of a positive relationship between the two variables, and delves into the mechanisms of this relationship. The findings suggest that the mechanisms from the theoretical framework can explain only part of the total correlation between skill-biased tariffs and growth. The paper attributes the remaining part of the correlation to the endogeneity of skill-biased tariffs, and in particular to the relationship between institutions and the skill-bias of tariffs: countries with good institutions tend to protect more skill-intensive industries.

In Nunn and Trefler (2010), three measures of the skill-bias of tariffs in the initial time period are used:[11] the correlation between the industry tariffs and the

[9]See Section A.2 in the Online Appendix for a description of the causal forest method. We consider the causal forest, and not the generic method developed by Chernozhukov et al. (2018b), as the latter requires a binary treatment variable.

[10]Table C.2 in the Online Appendix shows the results using default values of the parameters (which are reported in the notes of the Table). Due to the small sample size, we are unable to tune the parameters with cross-validation; thus, we perform sensitivity analysis varying the parameter values. The results, available upon request, are consistent with those reported in Table C.2.

[11]The initial time period is 1972 for 21 countries, 1980–83 for 30 countries and 1985-87 for

industry's skill-intensity, and two measures based on the difference between the log average tariffs in skill-intensive industries and log average tariffs in unskilled-intensive industries, which use different cut-off values for industry skill-intensity. In the country-level estimates, the outcome is log annual per capita GDP growth, and the regressions include a set of control variables.[12] The country-level regressions includes 63 observations.

For the industry-level estimates, the outcome variable is the average annual log change in industry output in each country, and the regressions include all the controls that appear in the country-level regressions, plus industry fixed effects. These regressions include 1004 data data points for 59 countries. An additional variable (the initial industry tariff) is included in some specifications to capture a potential mechanism: skill-biased tariffs can shift resources towards skill-intensive industries that generate positive externalities, thus leading to higher long-term growth. Thus, industries that have higher initial tariffs should have higher long-run output. If this channel can explain the effect of skill-bias on growth, the coefficient of the skill-bias of tariffs would decrease in size when this variable is included in the regression.

DML Analysis. We revisit the country and industry-level regressions reported in Tables 4 (columns 1, 2 and 4), Table 5 (columns 1, 2 and 4) and Table 6 (columns 1, 3 and 7) of Nunn and Trefler (2010). Further details on how the DML estimates are obtained and on the tuning parameter values are reported in Section B.2 of the Online Appendix.

Table 2 shows the results of the DML partially linear model using country-level data. The DML treatment effect estimates are considerably smaller than the original paper's across all ML methods and across the three different treatment variables. Moreover, the estimated effects are not statistically significant, except the coefficients estimated using the lasso, which are significant at the 10% level. Additionally, we report the DML results using the industry-level data set (Table C.3 and Table C.4 in the Online Appendix show the results with and without including the initial industry tariff respectively). Similarly to the country-level

12 countries. The end period is 2000 for most countries, except for 3 of them, for which data ends in 1996. See Nunn and Trefler (2010), Table 1 for a list of the countries included and the respective time periods.

[12]Further details on the regressions estimated by Nunn (2007) and on the control variables are described in Section B.2 of the Online Appendix.

Table 2: The Structure of Tariffs and Long-Term Growth: Country-level estimates

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) OLS
Panel A: Skill tariff correlation								
Skill tariff correlation	0.018	0.016	0.016	0.015	0.014	0.017	0.016	0.035
	(0.010)	(0.011)	(0.012)	(0.011)	(0.013)	(0.012)	(0.011)	(0.01)
Panel B: Tariff differential (low cut-off)								
Tariff differential (low cut-off)	0.009	0.006	0.007	0.008	0.013	0.009	0.008	0.016
	(0.005)	(0.005)	(0.005)	(0.005)	(0.008)	(0.006)	(0.005)	(0.005)
Panel C: Tariff differential (high cut-off)								
Tariff differential (high cut-off)	0.011	0.008	0.009	0.009	0.005	0.009	0.009	0.02
	(0.005)	(0.005)	(0.006)	(0.006)	(0.008)	(0.006)	(0.006)	(0.004)
Observations	63	63	63	63	63	63	63	63
Raw covariates	17	17	17	17	17	17	17	17

Notes: Analysis of Table 4 (columns 1, 2, 4) of Nunn and Trefler (2010) using DML. Column (8) reports the original paper estimates. Standard errors are reported in parentheses. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. The number of covariates does not include the treatment variable.

estimates, the industry-level estimates are not statistically significant across all methods, except for the boosting estimates, which are significant at the 10% level.

Overall, the DML results suggest that the correlation between skill-biased tariffs and long-term economic growth is not robust to controlling for an unknown function of the average tariff level, country characteristics, initial production structure, cohort and region fixed effects. Indeed, the fact that the DML estimates are insignificant points to the presence of nonlinear confounding effects that are not accurately captured by the OLS regressions.

It is worth noting here that the original paper attributes most of the correlation found between the treatment variables and long-term growth to the endogeneity of the skill-biased tariff variables, arising from the fact that skill-biased tariffs are more likely in countries with better institutions. Interestingly, in this example the country-level DML estimates are in line with the notion that the direct effect of the skill bias of tariffs is smaller than what is estimated by the OLS regressions. Finally, our results only concern the relationship between skill-biased tariffs and long-run economic growth, and not the relationship between skill-biased tariffs and institutions, or between institutions and long-run growth, which are examined in the original paper. Thus, our findings are consistent with the alternative mechanism described in Nunn and Trefler (2010), i.e., the existence of a causal relationship between institutions and economic growth.

Furthermore, we estimate the country-level regressions with the causal forest method. The ATEs obtained with the causal forest, reported in Table C.5 in the Online Appendix, are very similar to the DML estimates.[13]

3 Heterogeneous Treatment Effects

This section focuses on the analysis of HTE for the effect of Fox News on Republican voting (DellaVigna and Kaplan, 2007) using the causal forest method (Wager and Athey, 2018; Athey et al., 2019) and the effect of a teacher training intervention (Loyalka et al., 2019) using the generic machine learning method (Chernozhukov et al., 2018b).

[13]As in the first application, the values of the tuning parameters used are the default values, and they are reported in the notes of Table C.5. Results considering different values for the parameters are consistent with those reported and are available upon request.

3.1 The Effect of Fox News on the Republican Vote Share

Description of Original Analysis. In this section we revisit and further analyze the study by DellaVigna and Kaplan (2007). This paper examines the impact of media bias on voting outcomes. Specifically, it analyzes the impact of the entry of a conservative cable television channel, Fox News, on the Republican Party's vote share in the United States. To identify the causal effect of Fox News on voting, the authors investigate whether towns where Fox News became available between 1996 and 2000 experienced an increase in the vote share for the Republican Party in Presidential elections during the same time period. The estimation is performed on a data set at the town level, comprising information on 9256 towns.

We consider the main outcome variable, i.e. the change in the vote share for the Republican party between 1996 and 2000. The treatment variable is a dummy indicating whether Fox News had become available between 1996 and 2000. To capture potential confounders, a number of control variables are included in the regressions.[14]

DellaVigna and Kaplan (2007) find a positive effect of Fox News on the Republican vote share. Moreover, they explore heterogeneity along a selected set of town characteristics: the number of available cable channels, the share of urban population, and whether the town is in a swing or Republican district. They do this by adding to the regression interaction effects of these covariates with the treatment variable.[15]

Causal Forest Analysis. We perform the HTE analysis using the causal forest method. Exploring heterogeneous effects is important for this study, in order to understand whether there are town or district characteristics that act as effect modifiers. While the average effects are informative for the impact of Fox News on the whole sample, it is often the case that treatment effects are not homogeneous. It is possible that the effect of Fox News was concentrated in some areas only. Understanding better the characteristics of the areas which saw the strongest and weakest responses can shed light on the mechanisms. The aim of this exercise is two-fold. First, we take an agnostic view about the nature of heterogeneity, and we investigate whether there are town or district characteristics which are

[14] Further details on the regressions and on the control variables in DellaVigna and Kaplan (2007) are described in Section B.3 of the Online Appendix.

[15] The findings are reported in Table 6 of the original paper.

Table 3: Fox News - Causal Forest: Average treatment effects and test for heterogeneity

	(1) District dummies	(2) Cluster-robust
Fox News effect (ATE)	0.0065	0.0065
	(0.0016)	(.0027)
Fox News effect above median	0.013	0.0072
	(0.0024)	(0.0028)
Fox News effect below median	-0.0033	0.0044
	(0.0021)	(0.0048)
95% CI for the difference	(0.01009, 0.02255)	(-0.00806, 0.01374)
Observations	9256	9256

Notes: This table reports the estimated average treatment effect and a test for overall heterogeneity using the causal forest. Standard errors are reported in parentheses.

treatment effect modifiers. Second, we examine whether the HTE analysis from the original paper matches the results from the causal ML methods.

We focus on one of the two preferred specifications from the original paper: the one that includes district fixed effects. We present results for two versions of the causal forest, which account for district-level effects in different ways. In the first set of results, we include in the analysis dummy variables indicating the congressional district where the town is located. In the second set of results, we implement a cluster-robust version of the random forest developed by Athey and Wager (2019), where we treat each district as a separate cluster. The advantage of the cluster-robust causal forest is that it does not assume that clusters have an additive effect on the outcome. Further details on the clustered-robust causal forest and tuning parameter values used for the analysis are discussed in Section B.3 of the Online Appendix. Note that pointwise asymptotic normality for the causal forest is provided for cases where the number of covariates is relatively low, and the covariates are continuous. To circumvent this issue, we perform a robustness check using the approach implemented by Athey and Wager (2019), where we train a preliminary random forest on all covariates, after which we run a final random forest on a reduced number of features. The results are discussed in Section B.3 of the Online Appendix and are very similar to those presented in this section.

We begin by discussing the average treatment effect. The results are presented in Table 3. As in the original analysis, we find a positive and significant effect of

Fox News on the Republican vote share, both when including district dummies, and when implementing the clustered-robust causal forest; however, the standard error in the clustered forest is larger. Our results suggest that in towns where Fox News became available the Republican party obtained a higher vote share by 0.65 percentage points on average, compared to towns where Fox News was not available. The ATE estimates are similar to the original paper estimates, which range between 0.4 and 0.7 percentage points (reported in Table 4 of DellaVigna and Kaplan, 2007, columns 4-7).

Next, we want to assess whether the causal forest can recover heterogeneity of treatment effects. As pointed out in Athey and Wager (2019), we can group observations according to whether their estimated out-of-bag conditional average treatment effect (CATE) is above or below the median CATE, and we can estimate the average treatment effect separately for these two subgroups. These are reported in Table 3 as *Fox News effect above median* and *Fox News effect below median*. Note that these results should be interpreted with caution, as developing uniformly valid standard errors for the causal forest is still an open question and we do not adjust our inference for the fact that we use predictions that contain error to build our subgroups.[16] The difference between the two subgroup estimates is large when including district dummies, suggesting that there is potential for heterogeneity, and it is statistically significant, as indicated by the fact that the 95% confidence interval for the difference between the two estimates does not contain zero (see column 1 of Table 3). However, the same heuristic test for the clustered-robust forest does not detect significant heterogeneity in the treatment effect. This could indicate that heterogeneity in the model with district dummy variables is overstated, because the dummy variables cannot appropriately capture the district-specific effects. The cluster-robust causal forest offers a more flexible way to capture district-specific effects, and may be more suitable in this case.[17]

Although the results of the test for overall heterogeneity are mixed, it is still possible for heterogeneity to be present along some of the covariates. Hence, we

[16] To supplement our analysis, we implement an additional test for overall heterogeneity, inspired by the Best Linear Predictor method in Chernozhukov et al. (2018b). The results, reported in Table C.6 and discussed in Section B.3 of the Online Appendix, are in line with those obtained from the test in Table 3.

[17] Athey and Wager (2019) find a similar result in their application, when comparing the causal forest without clustering with the cluster-robust version.

Table 4: Fox News - Causal Forest: HTE analysis

	(1) CATE below median	(2) CATE above median	(3) *p*-value difference
		Panel A: District dummies	
Employment rate, diff. btw. 2000 and 1990	0.00928	0.00064	0.00656
	(0.00244)	(0.00203)	
Share high school degree 2000	0.00805	-7e-05	0.00884
	(0.00226)	(0.00213)	
Decile 10 in no. cable channels available	0.00877	-0.0044	6e-05
	(0.00192)	(0.00264)	
		Panel B: Cluster-robust	
Employment rate, diff. btw. 2000 and 1990	0.00938	2e-04	0.06885
	(0.00254)	(0.00436)	
Share high school degree 2000	0.00859	-0.00179	0.05296
	(0.00303)	(0.00442)	
Decile 10 in no. cable channels available	0.00857	-0.00495	0.02033
	(0.00289)	(0.00506)	

Notes: This table reports the effect of Fox News on the Republican vote share for towns with values below (column 1) and above (column 2) the median of each variable. Column 3 presents the *p*-value for the null of no difference between the estimates in columns 1 and 2. Standard errors are reported in parentheses.

investigate whether any of the included covariates are possible sources of heterogeneity. To do this, for each variable, we split the sample in two parts, based on whether the value of the covariate of interest is below and above the median, and we estimate the average treatment effect for the two subsamples. Table 4 reports the HTE results along the variables that appear to be significant determinants of heterogeneity in both specifications, while C.7 and C.8 in the Online Appendix report the results for the remaining variables. In addition, to gain further insight into which variables are more important for heterogeneity, we compute a measure of variable importance (Athey and Wager, 2019).[18] Tables C.9 and C.10 in the Online Appendix report the variable importance measure for the covariates included in the district dummy variable specification and for the clustered-robust forest, respectively. We note that for both specifications, the variable importance measure is decreasing smoothly and we do not observe any variable that clearly stands out in terms of importance.

Our results in Table 4 show that three variables appear to be significant determinants of heterogeneity (at least at the 10% level) in both specifications: the change in employment between 1990 and 2000, the share of the population with education level equal to high school degree, and the 10th decile in number of cable

[18] See Section B.3 of the Online Appendix for details on how this measure is constructed.

channels available. We observe that the effect of Fox News on Republican voting is stronger in towns that experienced a smaller increase in the employment rate between 1990 and 2000. This finding may relate to the phenomenon of economic voting, i.e. the fact that voters tend to reward incumbents during periods of economic prosperity (e.g., Fair, 1978; Kramer, 1971; Lewis-Beck and Stegmaier, 2000; Pissarides, 1980). Areas that experienced lower economic growth (and a smaller increase in employment) may have been more easily persuaded to vote Republican in 2000, since prior to the Presidential election of 2000 a Democratic President (Bill Clinton) had been in power for two consecutive mandates. Moreover, we observe a larger effect of Fox News in towns where the share of population with education level equal to high school degree is below median. We also find a larger positive effect of Fox News in towns where the 10th decile in the number of cable channels is below median, while the effect is negative and insignificant in towns where this variable is above median.[19]

Next, we investigate whether the findings regarding heterogeneity from the original paper are confirmed with the causal forest. DellaVigna and Kaplan (2007) found a larger effect of Fox News on the Republican vote share in towns with a smaller number of cable channels available when including district fixed effects. While we do not observe significant heterogeneity along this variable, our results for the 10th decile in the number of cable channels are in line with the findings of the original analysis, and hence suggest that the effect of Fox News diminishes in the presence of higher competition in cable channels. It is also interesting to note that the number of cable channels emerges as the variable with the highest importance score in both specifications, which further points to the importance of this variable for heterogeneity. When investigating heterogeneity along the political orientation of the district, we confirm the findings of DellaVigna and Kaplan (2007): we observe no significantly different effect for swing districts, and we obtained mixed results for Republican districts, as we find a significantly smaller effect of Fox News in Republican districts (at the 10% level) when including district dummies, but not with the cluster-robust forest.[20] However, in contrast

[19] The median value for the 10th decile in number of cable channels is zero; hence, towns with value of this variable above median correspond to towns that are in the top decile in terms of number of cable channels available.

[20] DellaVigna and Kaplan (2007) found mixed results for Republican districts in different specifications.

to the original analysis, we do not find a significant difference in the effect of Fox News in rural versus urban towns, despite this being the only heterogeneity result that is robust in all specifications in DellaVigna and Kaplan (2007).

In conclusion, our analysis of the HTE of Fox News on Republican voting confirms some of the findings from DellaVigna and Kaplan (2007), namely the presence of heterogeneity along the number of cable channels and no robust heterogeneous effects for districts with different political orientations, but as opposed to the original paper it does not show different effects for urban and rural areas. The analysis with the causal forest further uncovers additional heterogeneity that was previously unexplored, such as a larger effect in towns that experienced a smaller increase in the employment rate, and a larger effect in towns with a lower share of population with high school degree. Finally, including district dummy variables results in the causal forest detecting more heterogeneity in treatment effects compared to the cluster-robust version, both when implementing the overall heterogeneity test and when analysing the HTE in terms of individual covariates. However, the model with district dummy variables could overstate the heterogeneity compared to the cluster-robust forest if the district dummies do not appropriately capture the district-specific effects. This points to the need of a more careful treatment of the issue of clustered observations when employing causal forests for empirical applications (Athey and Wager, 2019).

3.2 The Effect of Teacher Training on Student Performance

Description of Original Analysis. We reanalyze a large-scale randomized experiment that investigates the effect of a teacher professional development (PD) program in China on student achievement and on other student and teacher outcomes. The experiment was first studied by Loyalka et al. (2019). Three hundred mathematics teachers, each employed in different schools across one province, took part in the intervention. The teachers were randomly assigned to one of the different treatment arms: PD only; PD plus a continuous follow-up with additional material and tasks for the trainees; PD plus an evaluation of the extent to which the teachers remembered the content of the training sessions; or no PD (control group). The PD intervention consisted of lectures and discussions.

Randomization was implemented at the school level, and in each school one teacher was nominated to participate in the intervention. The main results are obtained by estimating a cross-sectional regression, where the treatment variable is a dummy indicating the treatment arm that the school was assigned to. The data was collected at three points in time: at baseline, midline and endline. Outcomes are measured at midline, or endline, and the main outcome of interest is student math achievement.[21] The control variables include student characteristics, teacher characteristics and class size.[22]

The original paper finds no significant effect of the PD intervention on students' achievement after one academic year, neither for the PD intervention alone, nor for the PD combined with the follow up and/or the evaluation treatments. The authors also do not find any effect on other outcomes, such as teacher knowledge or student motivation. The lack of effectiveness of the program is attributed to several factors: the content was too theoretical, the PD was delivered passively, and teachers could face constraints in the implementation of the suggested practices in the schools. Furthermore, the paper analyzes heterogeneous treatment effects, by interacting the treatment variable with a number of student and teacher characteristics: student's household wealth, baseline achievement level, the amount of training the teacher has received prior to the intervention, student and teacher gender, whether the teacher has a college degree and whether the teacher majored in math. The findings suggest that the effect of the treatment on students' achievement can differ by teacher characteristic; however, no heterogeneous effects are found in terms of characteristics of students.

Generic ML Analysis. We extend the analysis of HTE conducted in the original paper, by implementing the generic machine learning method developed by Chernozhukov et al. (2018b). Exploring heterogeneous treatment effects is particularly relevant for this intervention, because a small and insignificant estimate for the ATE could hide significant heterogeneity. Our aim is to dig deeper into the analysis of heterogeneous treatment effects. First, we investigate whether there is significant heterogeneity in treatment effects; second, we analyze whether causal machine learning methods, by implementing a systematic search for heterogeneity across a large number of covariates, can offer additional insights about the

[21] As Loyalka et al. (2019) show similar results when estimating the impact of the intervention at midline or endline, we focus on the outcome variables measured at endline.

[22] Section B.4 of the Online Appendix describes the regressions and the control variables.

Table 5: Teacher Training - Generic Method: Best Linear Predictor

	(1) ATE (β_1)	(2) HET (β_2)
Estimate	0.002	0.651
90% Confidence Interval	(-0.068,0.072)	(0.312,0.990)
p-value	1.000	0.0003
Observations	10006	10006

Notes: The estimates are obtained using neural network to produce the proxy predictor *S(Z)*. The values reported correspond to the medians over 100 splits.

characteristics of those who benefited from the program and those who did not, compared to the traditional methods used in the original paper.

In our analysis, we focus on the main outcome of interest, i.e. student math achievement. Since the results in the original paper are consistently close to zero when comparing the three different treatment arms with the control group, we choose to only analyze one of the treatment arms, corresponding to the PD intervention plus the evaluation. The sample that we use includes 10,006 students in 201 schools. We follow Loyalka et al. (2019) and cluster standard errors at the school level. In addition to the full set of controls included in the original paper, we also add to our analysis other variables that could be treatment effect modifiers: the baseline values of a number of student-level variables, plus variables indicating teachers behaviour in the classroom, evaluated by students at baseline.[23]

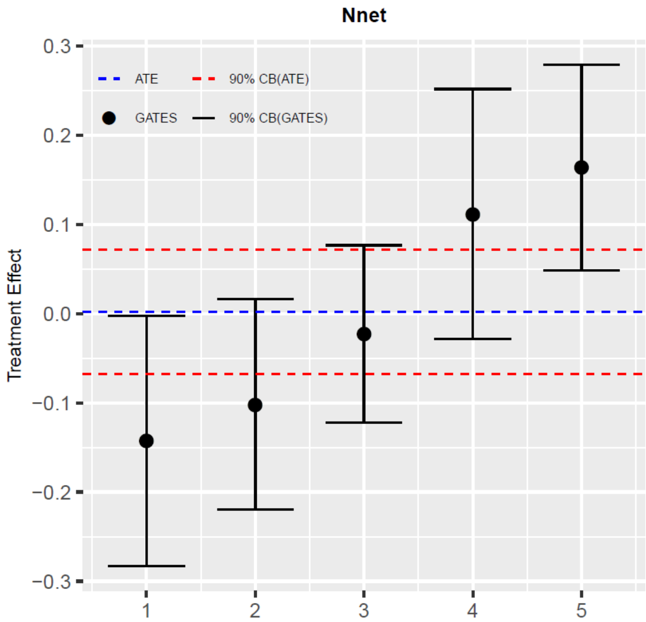
The generic method can be used in conjunction with a range of ML tools and Chernozhukov et al. (2018b) provide two measures (Best BLP and Best GATES) to compare the performance of the different ML methods used for the estimation of the proxy predictors. We consider the following methods: elastic net, neural network, and random forest. Based on the results of the Best BLP and Best GATES analysis, reported in Table C.13 of the Online Appendix, we choose to further work with the neural network.[24]

We first analyze whether overall heterogeneity in treatment effects can be de-

[23]These additional variables are described Section B.4 of the Online Appendix. In Loyalka et al. (2019), the baseline value of the outcome variable is included as a control. Hence, the baseline characteristics described above are not included in all regressions in the original analysis. However, we consider these characteristics as potential drivers of heterogeneity; therefore, we include the baseline values of all available variables in our heterogeneity analysis.

[24]Further details on the Best BLP and GATES measures and on the tuning parameters used in this analysis are discussed in Section B.4.

Figure 1: Teacher Training - Generic Method: GATES



Notes: The estimates are obtained using neural network to produce the proxy predictor *S(Z)*. The point estimates and 90% confidence intervals correspond to the medians over 100 splits.

tected. We present results for the best linear predictor (BLP) of the CATE in Table 5. In line with the original paper, the estimated ATE, given by the coefficient β_1, is small (the estimated impact of the PD is 0.002 standard deviations) and not significantly different from zero. The estimated β_2 is instead large and significantly different from zero, which indicates that there is heterogeneity in treatment effects. Next, we estimate the group average treatment effects (GATES). We split the sample into five groups, based on the quintiles of the ML proxy predictor *S(Z)*. This analysis reveals further insights into the extent of heterogeneity. Table C.14 of the Online Appendix reports the GATE in the top and bottom quintile and shows that the GATE in the top quintile is positive, whereas for the bottom quintile the estimated GATE is negative. Both estimates are statistically significant at the 10% level. The difference between the GATE for the top and the bottom quintile is significant, which confirms the presence of heterogeneity in treatment effects. Additionally, Figure 1 reports the GATES estimate and the 90% confidence interval for the five quintiles, as well as for the whole sample (the ATE is represented as a blue dashed line, and the confidence interval as two red dashed lines). Notice that for the three middle quintiles the effect of the teacher training

Table 6: Teacher Training - Generic Method: Classification Analysis

	(1) 20% most affected	(2) 20% least affected	(3) p-value for the difference
Teacher college degree	0.039	0.800	0.000
	(0.019,0.059)	(0.780,0.820)	
Teacher training hours	2.447	1.684	0.000
	(2.399,2.494)	(1.636,1.731)	
Teacher ranking	0.666	0.405	0.000
	(0.635,0.697)	(0.374,0.437)	
Student age	14.18	13.73	0.000
	(14.11,14.25)	(13.65,13.80)	
Teacher experience (years)	16.18	13.16	0.000
	(15.60,16.76)	(12.58,13.74)	
Student female	0.417	0.555	0.000
	(0.385,0.449)	(0.523,0.587)	
Teacher age	37.51	35.01	0.000
	(37.02,38.00)	(34.52,35.50)	
Student math score at baseline	-0.029	0.169	0.005
	(-0.088,0.031)	(0.110,0.229)	
Student baseline math anxiety	0.298	-0.219	0.000
	(0.236,0.360)	(-0.281,-0.157)	
Class size	52.87	64.37	0.000
	(51.82,53.93)	(63.32,65.43)	

Notes: This table shows the average value of the teacher and student characteristics for the most and least affected groups. The estimates are obtained using neural network to produce the proxy predictor *S(Z)*. 90% confidence intervals are reported in parenthesis. The variables *Student math score at baseline* and *Student baseline math anxiety* are normalized. The values reported correspond to the medians over 100 splits.

intervention is not significantly different from zero.

We then turn to analysing the possible sources of heterogeneity, by implementing the Classification Analysis (CLAN). Thus, we analyze further the top and botton quintile in terms of ATE, for which the effect of the PD intervention is positive and negative respectively. In particular, we compare the student and teacher characteristics in the two groups. As a large number of covariates is available, we focus on the ten covariates for which the correlation with the proxy predictor, $S(Z)$, is highest, reported in Table 6. Table C.15 in the Online Appendix shows the CLAN analysis for the remaining covariates. Table C.16 reports the correlation for each of the covariates with $S(Z)$.

We start by analyzing the *characteristics of the teachers* whose students belong to the least and most affected group. Interestingly, the variable indicating whether the teacher has a college degree or not is the variable that is most correlated with the proxy predictor, and it was the only one among the variables tested which was found to be a treatment effect modifier across all treatment arms in the original paper. The students in the top quintile are more likely to be taught by a teacher who does not have a college degree, compared to the students in the bottom quintile. This is consistent with the results from Loyalka et al. (2019), who found that the intervention has a negative effect on students whose teachers have a college degree, but a positive effect on students whose teachers are less qualified. Hence, the PD may help teachers who are less qualified, but, for more qualified teachers, the benefits of the intervention on their students do not outweigh the negative effect of the teachers being absent from the classroom in order to participate in the intervention. Whether or not the teacher majored in math is found to be a potential driver of heterogeneity with the generic method (the results are reported in Table C.15), whereas in the original paper it was not found to be significant when considering the effect of the PD plus evaluation, which we focus on.[25] The direction of the effect is consistent with what was found in the original analysis: the students in the top quintile are more likely to have been taught by a teacher who does not have a major in math, compared to the students in the bottom quintile. It is also interesting to note that the number of hours of training that the teacher received prior to the intervention, which is not found to be a determinant of heterogeneity in the original paper, is higher in the most affected group compared to the least affected group.[26] This may reflect the fact that teachers who have had more training in the past may be able to better implement the suggestions from the PD intervention. Table 6 shows that teacher rank, experience and age are higher in the most affected group compared to the least affected group. This is consistent with the existence of a similar mechanism: teachers who have more experience may be able to better implement the suggestions from the PD intervention. As the PD is mainly theoretical, having had other types of training, or having more experience, may be helpful for an effective implementation of the practices learned during the PD.

We then examine whether any of the *student characteristics* are potential drivers of heterogeneity. In contrast to the findings in Loyalka et al. (2019), who

[25] When considering the PD plus follow-up, the authors find a significant negative effect on the scores of students whose teachers majored in math relative to the scores of those whose teachers did not.

[26] The variable indicating teacher training hours previous to the intervention is a categorical variable, based on the terciles of the continuous variable. As the continuous variable is not included in the replication data set of the original paper, for our analysis we use this categorical variable, which takes values 1 to 3, where 3 is the top tercile in the number of training hours.

did not find heterogeneity in terms of student features, we find that students in the most affected group differ in terms of several characteristics compared to students in the least affected group. Among the most correlated with the heterogeneity score (listed in Table 6) are student age and gender: students in the most affected group are on average about half a year older than students in the least affected group, and the most affected group includes a larger share of male students. Additionally, students in the most affected group, on average, have a lower baseline math score, and tend to be more anxious about math. Thus, teacher PD could be more beneficial for weaker students, and for students who are more anxious about the subject. Finally, class size appears to be a possible determinant of heterogeneity: students who benefit more from the PD tend to be in smaller classes. This result suggests that in smaller classes it may be easier for teachers to implement some of the practices introduced during the PD training. For instance, Loyalka et al. (2019) mention having students work together in small groups as one of the techniques that were suggested in the PD; this technique is likely to be easier to implement in smaller classes.

In conclusion, our analysis confirms the presence of heterogeneous effects of the teacher PD intervention, and uncovers a rich set of potential determinants of heterogeneity. With the GATES analysis, we are able to show that the achievement of students belonging to the bottom quintile is negatively affected by the intervention, while the achievement of students in the top quintile is positively affected by the intervention. This confirms what was suggested by Loyalka et al. (2019): that there is a group of students who benefits from the intervention, and a group who does not. In addition, the GATES analysis shows that the effect is not significantly different from zero for the students belonging to the middle quintiles. With the CLAN analysis, we can obtain a clearer picture of the characteristics of the groups who benefit and who do not from the intervention, compared to the original HTE analysis. In line with Loyalka et al. (2019), we find that teacher characteristics such as having a college degree or having a major in math are potential determinants of heterogeneity. However, our study uncovers additional differences (that were not identified in the original paper) between the least and the most affected groups, in terms of both teacher and student characteristics, such as teacher's rank, experience, age and number of training hours, as well as student's gender, age, baseline math score, baseline math anxiety and class size.

4 Conclusion

Our main message is that appropriately combining predictive methods with causal questions adds value to traditional methods and should be more often explored in applied research. We argue that in each revisited study the researcher would have benefited from employing causal ML methods and would have gained additional insights not provided by standard causal inference tools.

We offer the following recommendations for applied researchers about the usefulness of causal ML methods.

1. Causal ML methods are useful in setting with many potential confounders relative to the sample size. With our revisited examples, we show the importance of taking into account all potentially relevant confounders at once, both linearly and nonlinearly. We revisit the most complete robustness checks of Djankov et al. (2010), and Nunn and Trefler (2010), considering all potential confounders linearly and nonlinearly, which would not be possible with traditional methods. Furthermore, our results from the Monte Carlo study suggest that, as the number of covariates used in the estimation increases relative to the sample size, the gains from using DML over OLS increase.

2. Causal ML methods are more suitable than traditional methods to flexibly capture the effect of covariates. As the true functional form is unknown, with flexible estimation we can better capture the effect of confounders. For instance, when revisiting the results of Djankov et al. (2010), we show suggestive evidence that there exist relevant nonlinear terms, which were not taken into account in the original analysis but are captured by the DML estimation. Moreover, we show with our MC simulations that in the presence of nonlinear confounders DML outperforms OLS.

3. We further recommend using causal ML methods in settings where the researcher does not have a lot of guidance from theory on which covariates should be included. This is because they implement a systematic model selection, rather than choosing an *ad hoc* specification, as we discuss when revisiting the results of Djankov et al. (2010). This argument is also important

when performing sensitivity analysis and robustness checks, as highlighted by our results when revisiting Nunn and Trefler (2010).

4. Finally, if the researcher is interested in HTE, causal machine learning methods can ensure that relevant heterogeneity and its determinants are not missed, or falsely discovered due to multiple hypothesis testing issues. For example, our analysis of the HTEs of the papers by DellaVigna and Kaplan (2007) and Loyalka et al. (2019) reveals potential determinants of heterogeneity that were not considered in the original analyses, which rely on traditional methods. Moreover, causal ML methods can be used to uncover heterogeneity ex-post, without being bound to explore HTE only for the specific subgroups indicated in the pre-analysis plan.

These advantages are particularly important in the context of observational studies, where causal ML methods improve the credibility of causal analysis by making the unconfoundedness assumption more plausible. However, even if the empirical study is a randomized control trial and controlling for confounding factors is not necessarily needed, the use of causal machine learning methods can improve efficiency and provide more precise estimates with lower standard errors and tighter confidence intervals.

References

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Athey, S. and Imbens, G. (2019). Machine learning methods economists should know about, arxiv.

Athey, S. and Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623.

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*.

Bertrand, M., Crépon, B., Marguerie, A., and Premand, P. (2017). Contemporaneous and post-program impacts of a public works program: Evidence from côte d'ivoire. Working Paper.

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

Carvalho, C., Feller, A., Murray, J., Woody, S., and Yeager, D. (2019). Assessing treatment effect variation in observational studies: Results from a data challenge. *Observational Studies*, 5(2):21–35.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018b). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Working Paper, National Bureau of Economic Research.

Colangelo, K. and Lee, Y.-Y. (2020). Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*.

Davis, J. and Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50.

Davis, J. M. and Heller, S. B. (2020). Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. *Review of Economics and Statistics*, 102(4):664–677.

DellaVigna, S. and Kaplan, E. (2007). The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., and Reif, J. (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, 109(12):4178–4219.

Djankov, S., Ganser, T., McLiesh, C., Ramalho, R., and Shleifer, A. (2010). The effect of corporate taxes on investment and entrepreneurship. *American Economic Journal: Macroeconomics*, 2(3):31–64.

Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., et al. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Fair, R. C. (1978). The effect of economic events on votes for president. *The Review of Economics and Statistics*, pages 159–173.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Grossman, G. M. and Helpman, E. (1991). *Innovation and growth in the global economy*. MIT press.

Hahn, P. R., Dorie, V., and Murray, J. S. (2019). Atlantic causal inference conference (acic) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Jacob, D. (2021). Cate meets ml-the conditional average treatment effect and machine learning. *arXiv preprint arXiv:2104.09935*.

Knaus, M., Lechner, M., and Strittmatter, A. (2018). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. Working Paper, CEPR Discussion Paper No. DP13402.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2020). Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, pages 0718–9615R1.

Kramer, G. H. (1971). Short-term fluctuations in us voting behavior, 1896–1964. *American Political Science Review*, 65(1):131–143.

Lewis-Beck, M. S. and Stegmaier, M. (2000). Economic determinants of electoral outcomes. *Annual Review of Political Science*, 3(1):183–219.

List, J. A., Shaikh, A. M., and Xu, Y. (2016). Multiple hypothesis testing in experimental economics. *Experimental Economics*, pages 1–21.

Loyalka, P., Popova, A., Li, G., and Shi, Z. (2019). Does teacher training actually work? evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics*, 11(3):128–54.

Nunn, N. (2007). Relationship-specificity, incomplete contracts, and the pattern of trade. *The Quarterly Journal of Economics*, 122(2):569–600.

Nunn, N. and Trefler, D. (2010). The structure of tariffs and long-term growth. *American Economic Journal: Macroeconomics*, 2(4):158–94.

Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2019). Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR.

Pissarides, C. A. (1980). British government popularity and economic performance. *The Economic Journal*, 90(359):569–581.

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2018). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *arXiv preprint arXiv:1712.09988*.

Strittmatter, A. (2019). What is the value added by using causal machine learning methods in a welfare experiment evaluation? Working Paper.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158.

Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in medicine*, 37(23):3309–3324.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Online Appendix

A Methodology

A.1 Double Machine Learning

The method is suitable in settings with a large number of covariates relative to the sample size (either because the number of raw covariates is large to begin with, or there is a large number of technical controls), where typical non-parametric kernel or spline methods break down.

The main model specification of the method, in the notation of Chernozhukov et al. (2018a), is the partially linear regression:

$$Y = D\theta_0 + g_0(X) + U \tag{1}$$

$$D = m_0(X) + V \tag{2}$$

where Y is the outcome, D is the treatment variable of interest, X is a (high-dimensional) vector of controls, and U and V are disturbances. Equation (1) is the main equation of interest and the parameter θ_0 is the treatment effect we would like to estimate. In this model, θ_0 quantifies the *average treatment effect*. The second equation is not of direct interest, but it keeps track of the dependence of the treatment on confounders. The covariates are related to the treatment through the function $m_0(X)$ and to the outcome variable through the function $g_0(X)$. While $m_0(X)$ and $g_0(X)$ can be nonlinear, the treatment variable, D, enters the model linearly (and additively). In observational studies, $m_0 \neq 0$, which means that the treatment assignment is not random, but depends on the covariates.

A first idea one might have for estimating θ_0 with ML methods would be to use a predictive-based ML approach and predict Y using D and X to obtain $D\hat{\theta}_0 + \hat{g}_0(X)$. This can be done for example by an iterative method that alternates between estimating g_0 with some ML method and θ_0 with OLS. While this 'naive' ML approach will have very good prediction performances, the iterative ML estimator will be heavily biased with a slower than $1/\sqrt{n}$ convergence rate.

The primary reason for this poor performance is the bias introduced by *regularization.* In order to optimize prediction and avoid overfitting the data with complex functional forms, ML methods use regularization and shrink the less important coefficients towards zero. This reduces overfitting by decreasing the variance of the estimator but at the same time introduces bias. The bias in estimating g_0 transfers to the parameter of interest θ_0. The issue is similar to the omitted variable bias.

To overcome regularization bias, Chernozhukov et al. (2017) propose 'double machine learning' i.e., solving two predictions problems instead of one. First, a ML model is fitted for m_0 in the treatment equation, and the effect of X is partialed out from D to get the residuals $\hat{V} = D - \hat{m}_0(X)$. Second, a ML method is fitted for the outcome equation and the residuals $\hat{W} = Y - \hat{l}_0(X)$ are obtained, where $l(X) = E[Y|X]$.[27] Finally, the residuals $\hat{W}$ are regressed on the residuals $\hat{V}$ to obtained the 'debiased' machine learning estimator, $\check{\theta}_0$. It can be shown that $\check{\theta}_0$ removes the effect of the regularization bias.[28]

However, $\check{\theta}_0$ is still subject to bias due to *overfitting*. For instance, when $\hat{g}_0$ is overfit, it will mistake noise for signal and thus it will pick up some of the noise U from the outcome equation. If U and V are correlated, the estimation error in $\hat{g}_0$ will be correlated with V. To break this correlation and avoid bias due to overfitting, one can rely on sample splitting. To this end, the data is partitioned into two subsamples: a main sample and an auxiliary sample. The ML models for the two nuisance functions m_0 and g_0 are fit on the auxiliary sample, while the residual on residual regression to obtain $\check{\theta}_0$ is fit on the main sample.

A drawback of sample splitting is that the estimator of the parameter of interest θ_0 is obtained using only the main sample, which can lead to loss of efficiency. However, one can switch the role of the main and auxiliary samples (procedure called *cross-fitting*) and average the results, which will lead to a more efficient estimator. In addition, one can perform a K-fold version of the cross-fitting procedure, where the size of each fold is n/K. Each sample partition or fold is successively taken as the main sample while the complement for each fold will be the auxiliary

[27] The nuisance functions can be estimated with a variety of ML methods such as: lasso, regression trees, random forest, boosting, neural networks, or hybrid methods.

[28] This is because the scaled estimation error, $\sqrt{n}(\check{\theta}_0 - \theta_0)$, contains now a term based on the product of two estimation errors (the estimation errors in $\hat{m}_0$ and in $\hat{l}_0$), which vanishes faster than the equivalent term obtained from using the naive estimator that depends on the estimation error in $\hat{g}_0$.

sample. One can take then the average of the estimates over the K-folds. To make the results robust to data partitioning, the splitting in folds procedure is performed S times, and the final DML estimator is the mean (or median) over the splits. The median version is more robust to outliers and this is the one we use in the applications.

A.2 Causal Forest

The causal forest method is an adaptation of the original random forest for prediction, introduced by Breiman (2001), to the problem of causal inference. In this section, we start by briefly presenting the general idea of standard regression trees used for prediction, after which we describe how causal trees and causal forest work.

The idea of *regression trees* is to partition (or split) the data into groups based on the values of the covariates. The groups that are eventually obtained are referred to as leaves. First, one starts with the whole data set as one group. Then, for each value of each covariate, the regression tree algorithm forms candidate splits, by placing all observations that have a covariate value that is lower than than the current value in the left leaf, and all observation for which their covariate value is greater than the current value in the right leaf. Among all these candidate splits, the one that is implemented is the one that minimizes an in-sample criterion function, such as the mean squared error (MSE) of the outcome variable within a leaf.[29] For each of the two new leaves, the algorithm repeats the procedure until a stopping rule[30] is reached, resulting in a tree-format partition of the data. Using the terminal leaves, when the purpose is prediction, the outcome variables of out-of-sample observations can be predicted by determining which terminal leaf a new observation belongs to, based on the values of the covariates, and assigning as its predicted outcome the mean of the outcomes in that leaf.

Next, we turn to the *causal forest* methods in Wager and Athey (2018) and Athey et al. (2019) which build on the *causal tree* method of Athey and Imbens (2016). For the causal tree, first, a percentage p from the sample N is drawn

[29]This mean squared error is computed as the sum of the squared differences between the outcomes of each unit within a leaf and the mean of these units in the leaf.

[30]The stopping criteria can be for example: a pre-specified maximum number of leaves, the iteration when the minimizing split gives a covariate over which the observations have been already split by, or the iteration when the proposed split does not decrease the mean squared error any further.

without replacement. Then, the subsample $n = p * N$ is further randomly split in half to form a training sample n_{tr} and an estimation sample n_e. Using only the training sample n_{tr}, for each value of each covariate candidate splits are formed and a regression tree as described above is constructed. The key difference in the causal case compared to the prediction case is the objective function that is optimized when determining the split to be implemented.

Due to the fundamental problem of causal inference, directly training machine learning methods on the difference $Y_i(1) - Y_i(0)$, i.e., the difference of the outcomes that observation i would have experienced with and without the treatment, is not possible, as we do not observe both outcomes for any individual unit. Thus, instead of minimizing an infeasible MSE, Athey and Imbens (2016) propose to maximize a criterion function that rewards a split that increases the variance of treatment effects across leaves and penalizes a split that increases within-leaf variance. The goal is to accurately estimate treatment effects within leaves, while preserving heterogeneity across leaves. The split is performed if it increases the criterion function, compared to no split. When no more splits can be done, the tree constructed based on the first subsample is defined.[31]

The subsequent step involves turning to the estimation sample n_e, and based on the covariates, sorting each observation in this sample into the same tree. Using only the estimation sample, the treatment effect in each leaf is computed as $\hat{\tau}_l = \bar{y}_{lt} - \bar{y}_{lc}$ i.e., the mean outcome difference between treated (t) and control (c) observations within a leaf (l). The final step consists in returning to the full sample of N observations, examining to which leaf each observation belongs based on the values of their covariates, and assigning that leaf's treatment effect as the predicted treatment effect of the observation. Given that estimates from a single tree can have a high variance, the whole algorithm described above is repeated for a number of B subsamples on which a number of B trees are obtained that eventually form a causal forest. The predicted treatment effect for each unit will be the average of predictions for that particular observation across the trees.

Notice that independent samples are used for: i) growing the tree (splitting the data), and ii) estimating treatment effects within each leaf of the tree. This prop-

[31] The causal forest of Athey et al. (2019) applies an approximated version of the splitting criterion of Athey and Imbens (2016) to get weights and runs a weighted partially linear regression in a final step (see Athey et al., 2019).

erty is called *honesty*. Honesty leads to two desirable characteristics: it reduces bias from overfitting, and it makes the inference valid, since the asymptotic properties of the treatment effect estimates are the same as if the structure of the tree had been exogenously given.[32] Further details regarding the tuning parameters of the causal forest are provided in Section B.3 of the Online Appendix.

A.3 Generic Machine Learning

A different causal ML approach for HTE is the generic machine learning method of Chernozhukov et al. (2018b). To make inference possible, the method does not focus directly on the HTEs, but on *features* of HTEs such as: the best linear predictor of the heterogeneous effects (BLP), the group average treatment effects (GATES) sorted by the groups induced by machine learning proxies, and the average characteristics of the units in the most and least affected groups, or classification analysis (CLAN). The generic machine learning method is thus useful for empirical work as: (1) it allows detection of heterogeneity in the treatment effect, (2) computes the treatment effect for different groups of observations (such as least affected or most affected groups), and (3) describes which covariates are correlated the most with the heterogeneity.

The approach is based on random splitting of the data into an auxiliary and a main sample. The two samples are approximately equal in size. Based on the auxiliary sample, a ML estimator, called proxy predictor, is constructed for the conditional average treatment effect (CATE). Any generic ML method can be used for this approximation (e.g., elastic net, random forest, neural network, etc.). The proxy predictors are possibly biased and consistency is not required. We simply take them as approximations and use them to estimate and make inference on features of the CATE. Based on the main sample and the proxy predictors, we can compute the estimates of interest: BLP, GATES and CLAN, and then make inference relying on many splits of the data in auxiliary and main samples.

We give a brief description on how the method works in practice. Let Y be the outcome of interest, D the binary treatment variable, and Z a vector

[32] Sample splitting, in general, can be inefficient as part of the data is not used. However, this loss of precision does not happen in the case of causal forests. This is because although no observation is allowed to be used within the same tree for both partitioning the covariate space and estimation, when the data is subsampled and the forest is obtained based on many trees, each individual unit will appear in both the training sample and the estimation sample of some tree.

of covariates. Define $b_0(Z) = E[Y(0)|Z]$, the baseline conditional average and $s_0(Z) = E[Y(1)|Z] - E[Y(0)|Z]$, the conditional average treatment effect (CATE). Using the auxiliary sample we obtain ML estimators (or proxy predictors) for the baseline conditional average and the conditional average treatment effect. As mentioned above, these are possibly biased predictors and consistency is not required. Then, for each unit in the main sample, we compute the predicted baseline effects, $B(Z)$ and the predicted treatment effects, $S(Z)$. Note that the predicted treatment effects, $S(Z)$, are obtained as the difference between the predictions for the treatment group model and the control group model. Following the notation from Chernozhukov et al. (2018b), the BLP parameters are obtained using the main sample, by estimating the following regression by weighted OLS, with weights $1/(p(Z)(1-p(Z))$:

$$Y = \alpha' X_1 + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S(Z) - \overline{S(Z)}) + \epsilon, \tag{3}$$

where $X_1 = [1, B(Z)]$, $p(Z) = P[D = 1|Z]$ is the propensity score, and $\overline{S(Z)}$ is the average of the predicted treatment effect estimates on the main sample. The control $B(Z)$ is included to improve efficiency. Note that the component $(D-p(Z))$ is part of the regressor $(D - p(Z))(S(Z) - \overline{S(Z)})$. Thus, it orthogonalizes this regressor to all other covariates that are functions of Z. The coefficient β_1 gives the average treatment effect, while β_2 quantifies how well the proxy predictor approximates the treatment heterogeneity. If β_2 is different from zero, it means that there exists heterogeneity in the treatment effects.

Once we obtain the predicted treatment effects, we can divide the observations from the main sample in groups: $G_1, G_2, \ldots, G_K$, based on their treatment effects. In our empirical applications, we choose $K = 5$, such that group G_1 contains the observations with the lowest 20% treatment effects and G_5 contains observations with the highest 20% treatment effects. Then, using again the main sample, we obtain the sorted group average treatment effects by estimating the weighted regression:

$$Y = \alpha' X_1 + \sum_{k=1}^{K} \gamma_k (D - p(Z)) \cdot 1(G_k) + \nu, \tag{4}$$

where $1(G_k)$ is an indicator function for whether an observation is in group k, and where the weights are the same as in (3). The parameters γ_k give the average effect in each group (GATES). Also, if the difference $\gamma_K - \gamma_1$ is significantly different from zero, we again have evidence for treatment effect heterogeneity between the most affected and least affected groups.

Lastly, we can analyze the properties or characteristics of the most affected and least affected groups, via Classification Analysis (CLAN). Let $g(Y, Z)$ be a vector of characteristics of an observation. We can compute average characteristics of the most affected and least affected group i.e., $\delta_1 = E[g(Y,Z)|G_1]$ and $\delta_K = E[g(Y,Z)|G_K]$, the parameters of interest being averages of variables directly observed. Similarly to GATES, we can compute and make inference on the difference $\delta_K - \delta_1$.

B Details on Revisited Studies and Implementation of Causal ML Methods

B.1 The Effect of Corporate Taxes on Investment and Entrepreneurship

Details on the Original Analysis. In Djankov et al. (2010), the baseline regression equation is the following:

$$y_c = \alpha + \beta taxes_c + \boldsymbol{X_c}\boldsymbol{\Gamma} + \epsilon_c,$$

where c is an index for country. Four different outcome variables are examined: investment as a percentage of GDP, FDI as a percentage of GDP, business density per 100 people, and the average entry rate (measured as percentage). Three separate measures of corporate taxes are considered. The first is the statutory corporate tax rates, which is the marginal tax rate on income a corporation has to pay assuming the highest tax bracket. The second is the actual first-year corporate income tax liability of a new company, relative to pre-tax earnings. The third is the tax rate which takes into account actual depreciation schedules going five years forward.

The term $\boldsymbol{X_c}$ denotes the control variables, aimed at capturing the effect of potential confounding factors. This is an observational study, in which tax rates are not randomly assigned across countries. It is likely that there will be factors which are correlated with both the treatment (corporate tax rates), and with the outcomes (measures of entrepreneurship and investment). To deal with this issue, the effect of corporate taxes on the outcomes is estimated by adding several control variables to the regressions. The first set of control variables are measures of other taxes: the sum of other taxes payable in the first year of operation, VAT tax, sales tax, and the highest national rate on personal income tax. The second set of covariates include the logarithm of the number of tax payments made (which is used as a measure of the burden of tax administration), an index of tax evasion, and the number of procedures to start a business. The third set of controls are institutional variables: a property rights index, an indicator of the rigidity of employment laws, a measure of a country's openness to trade, and the log of per capita GDP. The fourth set of covariates are measures of inflation: average inflation in the previous ten years, and seigniorage, which captures government reliance on printing money.

Details on the DML Analysis. The results are based on 100 splits and 2 folds. The point estimates are calculated as the median across splits, and the standard errors are adjusted for the variability across sample splits using the median method, see Chernozhukov et al. (2018a).

We use two hybrid ML methods in our analysis. Ensemble is a weighted average of estimates from lasso, boosting, random forest and neural networks, the weights being chosen to give the lowest average mean squared out-of-sample prediction error. Best chooses the best method for estimating the nuisance functions in terms of the average out-of-sample prediction performance among all the other methods.

The lasso estimates are based on ℓ_1-penalized regressions with the penalty parameter obtained through 10-fold cross-validation. As controls, for the lasso we consider the set of all raw covariates as well as first-order interactions. For the rest of the ML methods, we consider the set of raw covariates as controls. The regression tree method fits a CART (classification and regression tree) tree with a penalty parameter (that restricts the tree from overfitting and makes sure that only splits that are considered "worthy" are implemented) obtained with 10-fold cross validation. The random forest estimates are obtained using 1000 trees, while the Boosting estimates are obtained with 1000 boosted regression trees. For the boosting, the minimum number of observations in trees' terminal nodes is set to 1 and the bag.fraction parameter is set to 0.5, except for Panel D of Table 1, where it is increased to 0.8. For the neural networks we used 2 neurons and a decay parameter of 0.01; the activation function is set to the linear function.[33] We perform further sensitivity checks where we increase the number of layers to 2, 3, and 4, the number of neurons to 4, and we also change the activation function to the softplus function (a smooth approximation of the Rectified Linear Unit (ReLU) function), and to the Tanh function.

For the *analysis of nonlinear terms with lasso*, we examine the estimated nuisance functions for the outcome *average entry rate* and the treatment variable *first-year effective tax rate*. In our analysis, for the estimation of the two nuisance functions, the lasso selects among the simple covariates, and their two-way interactions.[34] It is interesting to note that a large number of interaction terms is selected. Figure C.1 depicts the seven largest interaction terms and their coefficients in the treatment nuisance function $\hat{m}(\cdot)$ and in the outcome nuisance function $\hat{g}(\cdot)$.The lasso coefficients are calculated as the median coefficients across splits. Among these, some appear in both nuisance functions (the coefficients of the common terms are depicted in purple in Figure C.1). A particular issue that appears with the lasso when the interest is on analyzing the interaction terms is worth mentioning here. Since the lasso implements regularization by shrinking the smallest coefficients to zero, it is possible that interaction terms are included in the regression, but the coefficients of the raw covariates forming the interactions are shrunk to zero. It is thus important to check whether the raw covariates forming these interactions also appear in the regression. If the coefficients on the raw covariates are shrunk, the coefficient of the 'pure' interaction terms might not be properly captured and the found interaction terms might actually reflect the effect of the raw covariates, diminishing the importance of our uncovered nonlinearities. Thus, when analyzing the relevance of the interaction terms, we are careful to only report the coefficients of the interactions for which both main effects are included in the lasso estimation. The lasso coefficients of all the raw covariates are reported in Table C.1 of the Online Appendix.

[33]In general, the activation function can be set to the linear function for regression problems (when the outcome is continuous) and to the logistic function for classification problems (when the outcome is categorical).

[34]For the lasso estimation, depending on the application, other nonlinear terms could be added, such as the squares of the covariates, or three-way interactions.

B.2 The Effect of Skill-Biased Tariff on Growth

Details on the Original Analysis. For the country-level results, Nunn and Trefler (2010) estimate the following regression equation:

$$ln\, y_{c1}/y_{c0} = \alpha + \beta_{SB} SB\tau_{c0} + X_{c0}\beta_X + \epsilon_c,$$

where $ln\, y_{c1}/y_{c0}$ is the log annual per capita GDP growth in country c between the beginning and the end of the time period considered, $SB\tau_{c0}$ is a measure of initial skill-bias of tariffs, and X_{c0} represents the controls. Three measures of the skill-bias of tariffs are used: the first is the correlation between the industry tariffs and the industry's skill-intensity, while the second and third are based on the difference between the log average tariffs in skill-intensive industries and log average tariffs in unskilled-intensive industries (the two measures differ in the choice of the cut-off value for industry skill-intensity, with the second using a lower cut-off than the third). The controls include: the log of the initial average level of tariffs in the country, three country characteristics measured at the initial period (the log of GDP per capita, the log of human capital, and the log of the ratio of investment-to-GDP), cohort fixed effects (to account for the fact that countries have different initial time periods), region fixed effects (accounting for 10 different regions), and two measures of initial production structure (the log of output in skill-intensive and in unskilled-intensive industries separately).

Additionally, Nunn and Trefler (2010) estimate the following regression equation, using industry-level data:

$$ln\, q_{ic1}/q_{ic0} = \beta_q ln q_{ic0} + \beta_\tau ln \tau_{ic0} + \beta_E ln \bar{\tau}_{c0} + \beta_{SB} SB\tau_{c0} + X_{c0}\beta_X + \alpha_i + \epsilon_{ic},$$

where lnq_{ic1}/q_{ic0} is the average annual log change in industry output in industry i and country c; lnq_{ic0} is the log of industry output in the initial period; τ_{ic0} is the log initial-period tariff; $\ln\, \bar{\tau}_{c0}$ is the average tariff, $SB\tau_{c0}$ is one of the three measures of skill-bias of tariffs, and α_i are industry fixed effects. The variable X_{c0} indicates the controls which are the same as in the country-level regressions.

The original results show a strong, positive correlation between skill-biased tariffs and long-term per capita income growth at the country level (Table 4 in Nunn and Trefler, 2010). The correlation is strong also between the skill bias of tariffs and industry output growth, with and without including the initial industry tariff in the regression (Tables 5 and 6 in Nunn and Trefler, 2010 respectively). The fact that the size of the coefficient of skill-biased tariffs remains large when adding the variable initial industry tariffs suggests that the mechanism highlighted in the model, i.e. skill-biased tariffs shifting resources towards skill-intensive industries, cannot fully account for the correlation between the treatment variable and long-term growth. Nunn and Trefler (2010) further show, with country-level regressions, that the model mechanism can explain up to one quarter of the total correlation between the skill bias of tariffs and long-term growth (Table 7 in the original paper). The paper then investigates other alternative mechanisms that can explain the independent effect of skill-biased tariffs on output growth in Sections V, VI and VII, in the original paper.

Details on the DML Analysis. As in the previous examples, the results are obtained with 100 splits and 2-fold cross-fitting. We report median estimates of the coefficients across splits, and standard errors are adjusted for the variability across sample splits using the median method.

The tuning choices are the same as in the previous two examples except for Neural Network in the country-level regressions where the estimates are obtained using 3 neurons and a decay parameter of 0.001.

B.3 The Effect of Fox News on the Republican Vote Share

Details on the Original Analysis. To produce the main results (see Table IV in the original paper), the authors estimate the following regression:

$$v^{R,Pres}_{k,j,2000} - v^{R,Pres}_{k,j,1996} = \beta d^{FOX}_{k,2000} + \boldsymbol{\Gamma_{2000} X_{k,2000}} + \boldsymbol{\Gamma_{00-90} X_{k,00-90}} + \boldsymbol{\Gamma_c C_{k,2000}} + \theta_j + \epsilon_{k,j},$$

where k denotes a town in a congressional district j. The dependent variable is the change in Republican vote share between the 1996 and the 2000 presidential elections. The treatment variable $d^{FOX}_{k,j,2000}$ is an indicator variable taking the value of 1 for towns where Fox News was available by the year 2000, and 0 otherwise. The regression includes demographic controls at the town level: total population, the employment rate, the share of African Americans and of Hispanics, the share of males, the share of the population with some college education, the share of college graduates, the share of high school graduates, the share of the town that is urban, the marriage rate, the unemployment rate, and average income. These controls are added both as levels in 2000 ($\boldsymbol{X_{k,2000}}$) and as changes between 1990 and 2000 ($\boldsymbol{X_{k,00-90}}$), and aim at capturing possible confounders that could be correlated with both the availability of Fox News and voting. In addition to the demographic controls, the regression includes a set of cable system features, denoted by $\boldsymbol{C_{k,2000}}$, which are potentially correlated with the treatment variable. These are deciles in the number of channels provided and in the number of potential subscribers. Finally, fixed effects (congressional district fixed effects or county fixed effects) denoted by θ_j, are added to capture trends in voting that might be common to a geographical area and also correlated with Fox News availability. In the original analysis, standard errors are clustered at the cable company level. The paper also tests whether Fox News increased voter turnout and the Republican vote share in the Senate election.

The results from the heterogeneity analysis of DellaVigna and Kaplan (2007) show a negative but insignificant effect for swing district. Additionally, the authors find that the effect of Fox News on the Republican vote share is significantly smaller in towns where the number of cable channels is higher, suggesting a negative impact of higher competition on the effect of Fox News. Moreover, the effect is found to be significantly larger in more urban areas and smaller in more Republican districts. Regarding the latter two findings, the authors point out that in rural areas and in Republican districts the Republican party tends to have a larger vote base to begin with, thus diminishing the share of voters that could potentially be convinced by Fox News. Out of the four effects, only the differential effect for urban population is significant in both main specifications (county and district fixed effects). The interaction of the treatment variable with the Republican district variable is only significant when including county fixed effects, but not when including district fixed effects, and the opposite is true for the interaction of the treatment with the number of cable channels. The authors also make a note that they find a smaller effect in the South, but this result is not reported in their paper, and we do not focus on it in our analysis.

Details on the Analysis with the Causal Forest. There are a number of parameters to be set in the causal forest algorithm such as the number of trees, the size of the subsample, and the minimum number of control and treatment units in each leaf. The *number of trees* is typically chosen as a trade-off between computation times and the test error rate. A larger number of trees reduces the Monte Carlo error due to subsampling, which means that the treatment effect predictions will vary less across different forests. A higher number of *minimum treatment and control units* will lead to bigger leaves and a less deep tree. This will predict less heterogeneity. A smaller number will increase the variance as the treatment effect will be estimated with too few observations in a given leaf. Setting a smaller *subsample size* will decrease the dependence across trees, but will increase the variance of each estimate in a tree. The *sizes of the training and estimation samples* are typically fixed to 50% of the drawn subsample. If there are reasons to allocate more observations to one or the other sample, these proportions can be changed. In the algorithm, there is also a standard parameter for the *number of covariates considered for a split*, when building a tree, within a forest.[35]

In our analysis, the tuning parameter values are optimised via cross-validation, except the number of trees which is set to 2000. We performed sensitivity analysis with different values for the number of trees (1000 and 5000). The results are available upon request.

Note that we perform the *orthogonalization* as discussed in Section 6.1.1 of Athey et al. (2019), which is especially useful in case of *observational studies* like the Fox News paper. This means that we estimate the propensity score and the marginal outcomes by training separate regression forests. We then compute the residual treatment and residual outcome and train a causal forest on these residuals (see Chernozhukov et al. (2017) for similar othogonalization ideas).

In the *cluster-robust* causal forest (Athey and Wager, 2019), when constructing the subsample on which the forest is trained, we do not directly draw observations, but clusters. In addition, in the final step, when constructing the predicted out-of-bag treatment effects, an observation is considered out-of-bag if its cluster was not drawn in the subsample.

The *variable importance* measure reported in Tables C.9 and C.10 takes into account the proportion of splits over all trees for a particular variable, weighted

[35]This makes random forests different from bagged trees. In bagged trees the number of predictors considered for a split is equal to the total number of covariates the researcher considers, while in random forests, the number of predictors is strictly less than this total number. The procedure 'decorrelates' the trees (as the trees will be less similar) and the aggregation of predictions across trees will have a lower variance.

by depth, and it is useful for describing which covariates influence the most the final estimates when employing the causal forest, as the interpretability of a single tree is lost in this case. Recall from the main text that in the causal forest splits are performed if they maximize a criterion function that rewards splits that increase the variance of the treatment effect across leaves, while penalizing splits that increases the variance within a leaf. Hence, higher values for this measure indicate higher importance in terms of heterogeneity of treatment effects.

In addition to the overall heterogeneity test presented in Table 3, we implement a second approach to test heterogeneity, based on the Best Linear Predictor method in Chernozhukov et al. (2018b) (see Section A.3). A positive and significant coefficient on the 'differential forest prediction' (β_2 in Section A.3) implies that there is heterogeneity. As noted by Athey and Wager (2019), asymptotic results for this approach are not currently available, and thus the results should be interpreted with this caveat in mind. The findings, presented in Table C.6, are consistent with those in Table 3: significant heterogeneity is detected for the specification with district dummies, but not for the cluster-robust forest.

As a robustness check, we also perform the analysis using a procedure presented in Athey and Wager (2019): we train a preliminary random forest on all covariates, after which we run a final random forest on the features that see the largest number of splits in the preliminary forest. We do this to address a potential concern with the causal forest: pointwise asymptotic normality is provided only for cases in which the number of covariates is relatively low. The overall heterogeneity test, reported in C.11, shows that the findings are unchanged when employing this procedure, both for the specification with district dummies and for the cluster-robust forest. The analysis of the driving features of heterogeneity, presented in Table C.12, also shows that most of the results obtained when using one forest are robust: the covariates that appear to be significant determinants of heterogeneity with both the district dummy specification and with the cluster-robust forest are the difference in the employment rate between 2000 and 1990, and the 10th decile in the number of cable channels available. However, we do not find significant heterogeneity in terms of the share of the population with education level equal to high school degree, although it is among the 9 most important variables in the specification with district dummies.

B.4 The Effect of Teacher Training on Student Performance

Details on the Original Analysis. In Loyalka et al. (2019), the main results are obtained by estimating the following regression equation:

$$Y_{i,j} = \alpha_0 + \alpha_1 D_j + X_{ij}\alpha + \tau_k + \epsilon_{i,j},$$

where $Y_{i,j}$ is the outcome, measured at midline or endline, for student i in school j; D_j is a dummy variable indicating the treatment assignment; the vector X_{ij} includes the control variables, measured at baseline; τ_k indicates the block fixed effects.[36] The main outcome of interest, student achievement, is measured with a 35-minutes mathematics test at endline. The full set of control variables includes students characteristics (age, gender, parent educational attainment, household wealth), class size, and teacher characteristics (gender, age, experience, education level, rank, a teacher certification dummy, and a dummy indicating whether the teacher majored in math).

The findings from the heterogeneity analysis suggest that the program has a small positive effect on achievement of students taught by less qualified teachers and a negative effect on students whose teachers are more qualified. In addition, some evidence of heterogeneity is found in terms of whether or not the teachers majored in math, with a negative effect on achievement for those students whose teachers did major in math (this effect is only found when comparing the PD plus follow up with the control group).

Details on the Generic ML Analysis. In addition to the full set of controls included in the original paper, we add to our analysis the following variables: the baseline values of a number of student-level variables (math self concept, math anxiety, intrinsic motivation for math, instrumental motivation for math, time spent each week studying math), plus a number of variables indicating teachers behaviour in the classroom, evaluated by students at baseline (instructional practices of teacher, teacher care, classroom management of teacher, teacher communication).

[36] The schools were randomized within blocks. A block is defined by the year of study the student is enrolled in (i.e. grades 7, 8, or 9), and by the two agencies that implemented the intervention. Hence, the total number of blocks is six.

The generic ML method takes into account two sources of uncertainty: estimation uncertainty, as the final estimates of interest are obtained conditional on the auxiliary sample, and splitting uncertainty, as the data is randomly split in many auxiliary and main samples. The point estimates are obtained as the median estimates over the different splits of the data. The confidence intervals are constructed by taking the medians of the lower and upper bounds over the random splits. Their nominal level is adjusted to 90% to account for the splitting uncertainty. In a similar way, the p-values are computed based on the median of many random conditional p-values, with nominal level adjusted again for splitting uncertainty.

The Best BLP and Best GATES measures are based on maximizing the correlation between the proxy predictor of the conditional average treatment effect, $S(Z)$, and the true conditional treatment effect, $s_0(Z)$ (see Chernozhukov et al., 2018b). Table C.13 shows that this correlation is the largest for the Neural Network. Therefore, we carry out the HTE analysis using the Neural Network.

The values of the tuning parameters were optimized via cross-validation for the Elastic Net and Neural Network. For the random forest they are set to default values to save on computation time. For the random forest, the number of trees is set to 2000 and the number of covariates considered for a split is set to $\lfloor K/3 \rfloor$, which is 8 in our case.

C Additional Tables and Figures

Table C.1: The Effect of Corporate Taxes on Entrepreneurship: Lasso Coefficients of Raw Covariates

	(1) Outcome: Average Entry Rate	(2) Treatment variable: First-year Effective Tax Rate
Log of number of tax payments	-0,402	0,017
Procedures to start a business	-0,006	0,001
Seigniorage 2004	-0,003	0
Other taxes	-0,001	0,003
Rigidity of employment	0	0
Average inflation (1995-2004)	0	0
PIT top marginal rate	0	0,01
IEF Property Right Index	0	0,001
VAT and sales tax	0	-0,005
Tax evasion (GCR)	0,009	-0,004
Log GDP pc 2003	0,011	-0,02
EFW Freedom to Trade Internationally Index	0,305	-0,619

Notes: The table shows the lasso coefficients of the raw covariates, obtained by estimating the nuisance functions $g(\cdot)$ (column 1) and $m(\cdot)$ (column 2). The lasso coefficients are calculated as the median over splits.

Table C.2: The Effect of Corporate Taxes on Investment and Entrepreneurship: Causal Forest estimates

	(1) Investment	(2) FDI	(3) Business density	(4) Average entry rate
Statutory corporate tax rate	-0.074	-0.131	-0.070	-0.148
	(0.066)	(0.085)	(0.054)	(0.056)
First-year effective tax rate	-0.158	-0.189	-0.119	-0.129
	(0.087)	(0.081)	(0.064)	(0.054)
Five-year effective tax rate	-0.222	-0.165	-0.112	-0.152
	(0.090)	(0.073)	(0.062)	(0.062)
Observations	61	61	60	50
Raw covariates	12	12	12	12

Notes: Analysis of Table 5D of Djankov et al. (2010) using the causal forest. The tuning parameters are set as follows: the number of trees is 2000; the sizes of the training and estimation sample are set to 50%; the minimum treatment and control units are 5; the number of covariates considered for a split is 12. Standard errors are reported in parentheses. The number of covariates does not include the treatment variable.

Table C.3: The Structure of Tariffs and Long-Term Growth: Industry-level estimates

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) OLS
				Panel A: Skill tariff correlation				
Skill tariff correlation	0.026	0.019	0.188	0.080	0.035	0.153	0.146	0.064
	(0.053)	(0.072)	(0.103)	(0.124)	(0.045)	(0.115)	(0.128)	(0.02)
				Panel B: Tariff differential (low cut-off)				
Tariff differential (low cut-off)	0.011	0.013	0.078	0.044	0.022	0.058	0.058	0.032
	(0.034)	(0.028)	(0.042)	(0.071)	(0.022)	(0.072)	(0.078)	(0.01)
				Panel C: Tariff differential (high cut-off)				
Tariff differential (high cut-off)	0.017	0.011	0.055	0.050	0.018	0.063	0.058	0.040
	(0.036)	(0.031)	(0.035)	(0.064)	(0.030)	(0.058)	(0.065)	(0.009)
Observations	1004	1004	1004	1004	1004	1004	1004	1004
Raw covariates	36	36	36	36	36	36	36	36

Notes: Analysis of Table 5 (columns 1, 2, 4) of Nunn and Trefler (2010) using DML. Column (8) reports the original paper estimates. Standard errors are reported in parentheses. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Standard errors adjusted for clustering at the country level are reported in column 8. The number of covariates does not include the treatment variable.

Table C.4: The Structure of Tariffs and Long-Term Growth: Industry-level estimates

	(1) Lasso	(2) Reg. Tree	(3) Boosting	(4) Forest	(5) Neural Net.	(6) Ensemble	(7) Best	(8) OLS
Panel A: Skill tariff correlation								
Skill tariff correlation	0.046	0.020	0.164	0.086	0.035	0.142	0.103	0.066
	(0.045)	(0.063)	(0.091)	(0.109)	(0.051)	(0.105)	(0.111)	(0.019)
Panel B: Tariff differential (low cut-off)								
Tariff differential (low cut-off)	0.026	0.015	0.069	0.048	0.022	0.073	0.055	0.033
	(0.024)	(0.033)	(0.037)	(0.061)	(0.026)	(0.059)	(0.059)	(0.01)
Panel C: Tariff differential (high cut-off)								
Tariff differential (high cut-off)	0.023	0.013	0.068	0.044	0.019	0.063	0.048	0.039
	(0.021)	(0.029)	(0.040)	(0.059)	(0.021)	(0.051)	(0.056)	(0.009)
Observations	1004	1004	1004	1004	1004	1004	1004	1004
Raw covariates	37	37	37	37	37	37	37	37

Notes: Analysis of Table 6 (columns 1, 3, 7) of Nunn and Trefler (2010) using DML. Column (8) reports the original paper estimates. Standard errors are reported in parentheses. Standard errors adjusted for variability across splits using the median method are reported for the DML estimates. Standard errors adjusted for clustering at the country level are reported in column 8. The number of covariates does not include the treatment variable.

Table C.5: The Structure of Tariffs and Long-Term Growth: Country-level estimates with Causal Forest

	(1) Skill tariff correlation	(2) Tariff differential (low cut-off)	(3) Tariff differential (high cut-off)
Skilled tariff	0.014	0.010	0.004
	(0.011)	(0.006)	(0.004)
Observations	63	63	63
Raw covariates	17	17	17

Notes: Analysis of Table 4 (columns 1, 2, 4) of Nunn and Trefler (2010) using the causal forest. The tuning parameters are set as follows: the number of trees is 2000; the sizes of the training and estimation sample are set to 50%; the minimum treatment and control units are 5; the number of covariates considered for a split is 17. Standard errors are reported in parentheses. The number of covariates does not include the treatment variable.

Table C.6: Fox News - Causal Forest: Best Linear Predictor

	(1) District dummies	(2) Cluster-robust
Mean forest prediction	1.168	1.117
	(0.316)	(0.518)
Differential forest prediction	2.52473	0.453
	(0.460)	(0.638)
Observations	9256	9256

Notes: This table reports the tests for the average treatment effect (mean forest prediction) and for heterogeneity (differential forest prediction), based on the approach Chernozhukov et al. (2018b), using the causal forest. Standard errors are reported in parenthesis.

Table C.7: Fox News - Causal Forest: HTE analysis with district dummies

	(1) CATE below median	(2) CATE above median	(3) p-value for the difference
Population, diff. btw. 2000 and 1990	0.00413 (0.00242)	0.00806 (0.00189)	0.20027
Share with high school degree, diff. btw. 2000 and 1990	0.0086 (0.00199)	0.0029 (0.0027)	0.08938
Share with some college, diff. btw. 2000 and 1990	0.00736 (0.00207)	0.0039 (0.00227)	0.26069
Share with college degree, diff. btw. 2000 and 1990	0.00757 (0.00272)	0.00582 (0.00191)	0.59872
Share male, diff. btw. 2000 and 1990	0.00949 (0.00222)	0.0035 (0.00231)	0.06126
Share African American, diff. btw. 2000 and 1990	0.00629 (0.00243)	0.00666 (0.002)	0.90674
Share Hispanic, diff. btw. 2000 and 1990	0.00428 (0.00238)	0.00737 (0.00208)	0.32866
Unemployment rate, diff. btw. 2000 and 1990	0.00366 (0.00238)	0.00866 (0.00224)	0.12612
Married, diff. btw. 2000 and 1990	0.00698 (0.00202)	0.00562 (0.00257)	0.67592
Median income, diff. btw. 2000 and 1990	0.00628 (0.00224)	0.00653 (0.0023)	0.93661
Share urban, diff. btw. 2000 and 1990	0.00517 (0.00203)	0.00945 (0.0025)	0.18368
Population 2000	0.00492 (0.00252)	0.00662 (0.00164)	0.57185
Share with some college 2000	0.00328 (0.00204)	0.00964 (0.00249)	0.04809
Share with college degree 2000	0.00556 (0.00253)	0.00679 (0.00185)	0.6946
Share male 2000	0.0055 (0.00194)	0.00976 (0.00277)	0.20794
Share African American 2000	0.0025 (0.00271)	0.00739 (0.00172)	0.12759
Share Hispanic 2000	0.00136 (0.00225)	0.00799 (0.00217)	0.03386
Employment rate 2000	0.00557 (0.00232)	0.00771 (0.00215)	0.50069
Unemployment rate 2000	0.00541 (0.00214)	0.00741 (0.00235)	0.52906
Share married 2000	0.00683 (0.00228)	0.00585 (0.00229)	0.76121
Median income 2000	0.00501 (0.00218)	0.00712 (0.00223)	0.50006
Share urban 2000	0.00441 (0.0024)	0.00673 (0.0019)	0.44815
No. potential cable subscribers 2000	0.00818 (0.00238)	0.00594 (0.00169)	0.44436
Decile 1 in no. potential cable subscribers	0.00661 (0.0016)	-0.00787 (0.01626)	0.37539
Decile 2 in no. potential cable subscribers	0.00664 (0.00165)	0.00084 (0.00861)	0.50799
Decile 3 in no. potential cable subscribers	0.00612 (0.00151)	0.0171 (0.0065)	0.0999
Decile 4 in no. potential cable subscribers	0.00634 (0.0017)	0.0077 (0.00393)	0.75084
Decile 5 in no. potential cable subscribers	0.00667 (0.00174)	0.00357 (0.00371)	0.44915
Decile 6 in no. potential cable subscribers	0.00669 (0.00171)	0.00471 (0.00463)	0.68762
Decile 7 in no. potential cable subscribers	0.00668 (0.0017)	0.00531 (0.00492)	0.79269
Decile 8 in no. potential cable subscribers	0.00758 (0.00168)	-0.00131 (0.00405)	0.04239
Decile 9 in no. potential cable subscribers	0.0071 (0.00167)	0.00226 (0.00317)	0.17685
Decile 10 in no. potential cable subscribers	0.0045 (0.00207)	0.01139 (0.00188)	0.01393
No. cable channels available 2000	0.00816 (0.00684)	0.0065 (0.00148)	0.812
Decile 1 in no. cable channels available	0.00645 (0.0017)	0.00149 (0.02648)	0.85167
Decile 2 in no. cable channels available	0.00655 (0.00153)	0.01884 (0.0383)	0.74845
Decile 3 in no. cable channels available	0.00657 (0.00161)	0.00758 (0.01372)	0.94203
Decile 4 in no. cable channels available	0.00747 (0.00155)	-0.0101 (0.01332)	0.18996
Decile 5 in no. cable channels available	0.00553 (0.00158)	0.02402 (0.01216)	0.13149
Decile 6 in no. cable channels available	0.00569 (0.00164)	0.01323 (0.00648)	0.25923
Decile 7 in no. cable channels available	0.00585 (0.00192)	0.00953 (0.00253)	0.24565
Decile 8 in no. cable channels available	0.0068 (0.00178)	0.00355 (0.00398)	0.45524
Decile 9 in no. cable channels available	0.00576 (0.00181)	0.01239 (0.00288)	0.05169
Swing district	0.00685 (0.00201)	0.00602 (0.00272)	0.80599
Republican district	0.00693 (0.00187)	0.00084 (0.00264)	0.06021

Notes: The table reports the effect of Fox News on the Republican vote share for towns with values below (column 1) and above (column 2) the median of each variable. Column 3 presents the p-value for the null of no difference between the estimates in columns 1 and 2. Standard errors are reported in parentheses. The estimates are obtained from the causal forest that includes district dummy variables. As we are not interested in exploring heterogeneity along the congressional districts, the HTE results for district dummy variables are omitted from the table.

Table C.8: Fox News - Causal Forest: HTE analysis with cluster-robust causal forest

	(1) CATE below median	(2) CATE above median	(3) p-value for the difference
Population, diff. btw. 2000 and 1990	0.00357 (0.00398)	0.00829 (0.00299)	0.34201
Share with high school degree, diff. btw. 2000 and 1990	0.0088 (0.00311)	0.00225 (0.00323)	0.14407
Share with some college, diff. btw. 2000 and 1990	0.00809 (0.00281)	0.00194 (0.00431)	0.23202
Share with college degree, diff. btw. 2000 and 1990	0.00709 (0.00317)	0.00604 (0.00329)	0.8194
Share male, diff. btw. 2000 and 1990	0.00975 (0.00356)	0.00308 (0.00268)	0.13407
Share African American, diff. btw. 2000 and 1990	0.00547 (0.00346)	0.007 (0.00298)	0.7364
Share Hispanic, diff. btw. 2000 and 1990	0.00369 (0.00383)	0.00755 (0.00286)	0.41946
Unemployment rate, diff. btw. 2000 and 1990	0.00328 (0.00304)	0.00872 (0.00308)	0.20834
Married, diff. btw. 2000 and 1990	0.00622 (0.00327)	0.00639 (0.00339)	0.97002
Median income, diff. btw. 2000 and 1990	0.0065 (0.00354)	0.00609 (0.00282)	0.92735
Share urban, diff. btw. 2000 and 1990	0.00527 (0.00273)	0.00881 (0.00372)	0.44257
Population 2000	0.00577 (0.00398)	0.00636 (0.0027)	0.9022
Share with some college 2000	0.00532 (0.00321)	0.00785 (0.00376)	0.60916
Share with college degree 2000	0.00545 (0.00296)	0.00672 (0.00318)	0.76975
Share male 2000	0.00459 (0.00259)	0.01138 (0.00529)	0.24942
Share African American 2000	0.00198 (0.00518)	0.00731 (0.00265)	0.35943
Share Hispanic 2000	0.00071 (0.00378)	0.00825 (0.00314)	0.1245
Employment rate 2000	0.0043 (0.00293)	0.00892 (0.00416)	0.36452
Unemployment rate 2000	0.00539 (0.0027)	0.00728 (0.0035)	0.66907
Share married 2000	0.00684 (0.00278)	0.00561 (0.00355)	0.78466
Median income 2000	0.00546 (0.00381)	0.00648 (0.00272)	0.82677
Share urban 2000	0.00534 (0.00404)	0.00647 (0.00276)	0.81683
No. potential cable subscribers 2000	0.00744 (0.00616)	0.00587 (0.00285)	0.81685
Decile 1 in no. potential cable subscribers	0.00653 (0.00264)	-0.00486 (0.0162)	0.48767
Decile 2 in no. potential cable subscribers	0.00655 (0.00268)	0.00209 (0.0116)	0.70797
Decile 3 in no. potential cable subscribers	0.00594 (0.00258)	0.01893 (0.0111)	0.25437
Decile 4 in no. potential cable subscribers	0.00628 (0.00256)	0.00734 (0.00928)	0.91234
Decile 5 in no. potential cable subscribers	0.00677 (0.00273)	0.00051 (0.00724)	0.4189
Decile 6 in no. potential cable subscribers	0.00691 (0.00278)	0.00113 (0.00592)	0.37691
Decile 7 in no. potential cable subscribers	0.00685 (0.00241)	0.00351 (0.01161)	0.77827
Decile 8 in no. potential cable subscribers	0.00741 (0.00283)	-0.00051 (0.004)	0.10608
Decile 9 in no. potential cable subscribers	0.00683 (0.00294)	0.00274 (0.00409)	0.41683
Decile 10 in no. potential cable subscribers	0.004 (0.00384)	0.01147 (0.00369)	0.16066
No. cable channels available 2000	0.00562 (0.00773)	0.00678 (0.00255)	0.88643
Decile 1 in no. cable channels available	0.00646 (0.00274)	-0.00726 (0.03324)	0.68079
Decile 2 in no. cable channels available	0.00644 (0.00265)	0.01768 (0.01293)	0.39453
Decile 3 in no. cable channels available	0.00672 (0.00267)	0.00203 (0.01179)	0.69811
Decile 4 in no. cable channels available	0.00816 (0.00263)	-0.01645 (0.01506)	0.10755
Decile 5 in no. cable channels available	0.00484 (0.00251)	0.02998 (0.01806)	0.16774
Decile 6 in no. cable channels available	0.00537 (0.00289)	0.0133 (0.0045)	0.13846
Decile 7 in no. cable channels available	0.00602 (0.00304)	0.00867 (0.0042)	0.60869
Decile 8 in no. cable channels available	0.00675 (0.00291)	0.00327 (0.00503)	0.54915
Decile 9 in no. cable channels available	0.00543 (0.0027)	0.0139 (0.00631)	0.21689
Swing district	0.00634 (0.00308)	0.00736 (0.00515)	0.86489
Republican district	0.00665 (0.00286)	0.00079 (0.00604)	0.38064

Notes: The table reports the effect of Fox News on the Republican vote share for towns with values below (column 1) and above (column 2) the median of each variable. Column 3 presents the p-value for the null of no difference between the estimates in columns 1 and 2. Standard errors are reported in parentheses. The estimates are obtained from the cluster-robust causal forest.

Table C.9: Fox News - Causal Forest: Variable importance with district dummies

Variable	Importance (%)
No. cable channels available 2000	6.52
No. potential cable subscribers 2000	5.23
Share employed, diff. btw. 2000 and 1990	4.9
Share African American 2000	4.74
Share married 2000	4.39
Unemployment rate, diff. btw. 2000 and 1990	4.22
Decile 10 in no. cable channels 2000	4.16
Unemployment rate 2000	3.67
Share with high school degree, diff. btw. 2000 and 1990	3.56
Share with some college 2000	3.55
Population, diff btw. 2000 and 1990	3.41
Share male, diff btw. 2000 and 1990	3.38
Share Hispanic, diff btw. 2000 and 1990	3.27
Median income, diff btw. 2000 and 1990	3.25
Median income 2000	3.22
Share Hispanic 2000	3.21
Share married, diff btw. 2000 and 1990	3.07
Share African American, diff btw. 2000 and 1990	3.02
Population 2000	3.01
Employment rate 2000	2.72
Share with some college, diff btw. 2000 and 1990	2.67
Share male 2000	2.55
Share with college degree 2000	2.49
Share with college degree, diff btw. 2000 and 1990	2.23
Share with high school 2000	2.14
Decile 10 in no. potential cable subscribers	2.08
Share urban population, diff btw. 2000 and 1990	1.9
Decile 7 in no. cable channels available	1.75
Decile 9 in no. cable channels available	1.54
Share of urban population 2000	1.4
Republican district	0.78
Decile 8 in no. cable channels available	0.75
Swing district	0.74
Decile 9 in no. potential cable subscribers	0.36
Decile 8 in no. potential cable subscribers	0.07
Decile 7 in no. potential cable subscribers	0.02
Decile 6 in no. cable channels available	0.01

Notes: The table reports the importance of each variable obtained from the causal forest with district dummies. Variables with importance lower than 0.01% are omitted.

Table C.10: Fox News - Causal Forest: Variable importance in cluster-robust causal forest

Variable	Importance (%)
No. cable channels available 2000	10.4
No. potential cable subscribers 2000	8.22
Share with some college 2000	4.79
Unemployment rate, diff. btw. 2000 and 1990	4.35
Decile 9 in no. cable channels	4.16
Decile 10 in no. cable channels	4.09
Employment rate, diff. btw. 2000 and 1990	3.86
Share African American 2000	3.59
Median income 2000	3.39
Population, diff. btw. 2000 and 1990	3.31
Median income, diff. btw. 2000 and 1990	3.1
Share married 2000	2.9
Share male, diff. btw. 2000 and 1990	2.88
Decile 7 in no. cable channels	2.84
Unemployment rate 2000	2.56
Share African American, diff. btw. 2000 and 1990	2.55
Share Hispanic, diff. btw. 2000 and 1990	2.5
Share married, diff. btw. 2000 and 1990	2.4
Share Hispanic 2000	2.38
Share with high school degree, diff. btw. 2000 and 1990	2.28
Share urban, diff. btw. 2000 and 1990	2.23
Share male 2000	2.2
Decile 10 in no. potential cable subscribers	2.19
Population 2000	2.14
Share with some college, diff. btw. 2000 and 1990	2.08
Share with college degree 2000	1.98
Employment rate 2000	1.79
Share with college degree, diff. btw. 2000 and 1990	1.58
Share with high school degree 2000	1.57
Share urban 2000	1.55
Republican district	1.37
Swing district	0.98
Decile 8 in no. cable channels	0.97
Decile 9 in no. potential cable subscribers	0.51
Decile 8 in no. potential cable subscribers	0.21
Decile 7 in no. potential cable subscribers	0.06
Decile 6 in no. cable channels	0.03
Decile 6 in no. potential cable subscribers	0.02

Notes: The table reports the importance of each variable obtained from the cluster-robust causal forest. Variables with importance lower than 0.01% are omitted.

Table C.11: Fox News - Causal Forest: Average treatment effects and test for heterogeneity, Double Forest

	(1) District dummies	(2) Cluster-robust
Fox News effect (ATE)	0.0064	0.0068
	(0.0016)	(0.0027)
Fox News effect above median	0.013	0.0074
	(0.0024)	(0.0028)
Fox News effect below median	-0.0044	0.0042
	(0.0021)	(0.0051)
95% CI for the difference	(0.0111, 0.0236)	(-0.0083, 0.0146)
Observations	9256	9256

Notes: This table reports the estimated average treatment effect and a test for overall heterogeneity using the causal forest. Standard errors are reported in parentheses.

Table C.12: Fox News - Causal Forest: HTE analysis, Double Forest

	(1) CATE below median	(2) CATE above median	(3) p-value difference
	Panel A: District dummies		
No. cable channels 2000	0.00883 (0.00683)	0.00639 (0.00148)	0.72696
No. potential cable subscribers	0.00757 (0.00239)	0.00593 (0.00168)	0.57511
Empl. rate, diff. btw. 2000 and 1990	0.00924 (0.00244)	0.00021 (0.00204)	0.00448
Share African American 2000	0.00186 (0.00271)	0.00739 (0.00173)	0.08565
Share married 2000	0.00683 (0.00229)	0.0055 (0.00229)	0.68062
Unempl. rate, diff. btw. 2000 and 1990	0.00334 (0.00238)	0.00872 (0.00225)	0.10075
Decile 10 in no. cable channels	0.00857 (0.00192)	-0.00475 (0.00264)	5e-05
Unempl. rate 2000	0.005 (0.00215)	0.00753 (0.00235)	0.42731
Share high school degree, diff. btw. 2000 and 1990	0.00839 (0.00202)	0.00307 (0.00268)	0.113
	Panel B: Cluster-robust		
No. cable channels 2000	0.00696 (0.00834)	0.00681 (0.00255)	0.98641
No. potential cable subscribers 2000	0.00769 (0.00636)-	0.00614 (0.00293)	0.82479
Share with some college 2000	0.00566 (0.00341)	0.00815 (0.00385)	0.62783
Unempl. rate, diff. btw. 2000 and 1990	0.00332 (0.00303)	0.00918 (0.0032)	0.18415
Decile 9 in no. cable channels	0.00571 (0.0028)	0.01411 (0.00633)	0.22478
Decile 10 in no. cable channels	0.00879 (0.00302)	-0.00426 (0.00516)	0.02902
Empl. rate, diff. btw. 2000 and 1990	0.0099 (0.0026)	-0.00017 (0.00459)	0.05626
Share African American 2000	0.00158 (0.00526)	0.00771 (0.00274)	0.30138
Median Income 2000	0.00574 (0.00403)	0.00672 (0.00268)	0.84033

Notes: This table reports the effect of Fox News on the Republican vote share for towns with values below (column 1) and above (column 2) the median of each variable. Column 3 presents the p-value for the null of no difference between the estimates in columns 1 and 2. Standard errors are reported in parentheses.

Table C.13: Teacher Training - Generic Method: Comparison of ML methods

	(1) Elastic net	(2) Neural network	(3) Random forest
Best BLP	0.012	0.014	0.011
Best GATES	0.115	0.121	0.099

Notes: The table compares the performance of the three ML methods used to produce the proxy predictors. The performance measures Best BLP and Best GATES are computed as medians over 100 splits.

Table C.14: Teacher Training - Generic Method: GATES of most and least affected groups

	(1) 20% most affected	(2) 20% least affected	(3) Difference
Effect of teacher training on student achievement	0.164	-0.179	0.365
90% Confidence Interval	(0.048,0.279)	(-0.301,-0.058)	(0.198,0.533)
p-value	0.011	0.092	0.001

Notes: The estimates are obtained using neural network to produce the proxy predictor *S(Z)*. The values reported correspond to the medians over 100 splits.

Table C.15: Teacher Training - Generic Method: Classification Analysis

	(1) 20% most affected	(2) 20% least affected	(3) p-value for the difference
Baseline instructional practices of teacher	0.211	0.053	0.000
	(0.149,0.272)	(-0.009,0.114)	
Teacher's baseline classroom management	0.065	0.074	0.000
	(0.003,0.127)	(0.012,0.136)	
Teacher gender	0.529	0.536	1.000
	(0.497,0.562)	(0.504,0.568)	
Teacher certification dummy	0.970	1.000	0.000
	(0.962,0.977)	(0.992,1.008)	
Student's baseline instrumental motivation for math	-0.111	0.224	0.000
	(-0.173,-0.050)	(0.162,0.285)	
Student's baseline time spent each week studying math	-0.073	0.204	0.000
	(-0.142,-0.004)	(0.135,0.273)	
Student's baseline math self-concept	-0.380	0.317	0.000
	(-0.441,-0.319)	(0.256,0.378)	
Teacher majored in math	0.309	0.507	0.000
	(0.278,0.340)	(0.475,0.538)	
Mother education level	0.570	0.425	0.000
	(0.538,0.602)	(0.393,0.457)	
Teacher's baseline communication	-0.088	0.251	0.000
	(-0.152,-0.025)	(0.187,0.314)	
Student's baseline intrinsic motivation for math	-0.294	0.299	0.000
	(-0.356,-0.232)	(0.237,0.362)	
Baseline teacher care	-0.165	0.311	0.000
	(-0.228,-0.103)	(0.249,0.374)	
Household asset index	-0.421	0.240	0.000
	(-0.491,-0.351)	(0.170,0.310)	
Father education level	0.583	0.589	1.000
	(0.551,0.614)	(0.557,0.621)	

Notes: The table shows the average value of the teacher and student characteristics for the most and least affected groups. The estimates are obtained using neural network to produce the proxy predictor *S(Z)*. Confidence intervals with 90% nominal level are reported in parenthesis. All variables, except *Teacher gender, Teacher certification dummy, Teacher majored in math, Mother education level* and *Father education level* are normalized. The values reported correspond to the medians over 100 splits.

Table C.16: Teacher Training - Generic Method: Correlation of the covariates with $S(Z)$

Variable	Correlation
Teacher college degree	-0.237
Teacher training hours	0.125
Teacher ranking	0.116
Student age	0.111
Teacher experience (years)	0.101
Student female	-0.094
Teacher age	0.089
Math score at baseline (normalized)	0.075
Student baseline math anxiety	0.063
Class size	-0.060
Baseline instructional practices of teacher	0.053
Teacher's baseline classroom management	0.051
Teacher gender	-0.045
Teacher certification dummy	0.036
Student's baseline instrumental motivation for math	0.025
Student's baseline time spent each week studying math	0.022
Student's baseline math self-concept	-0.021
Teacher majored in math	-0.016
Mother education level	0.009
Teacher's baseline communication	-0.008
Student's baseline intrinsic motivation for math	0.008
Baseline teacher care	-0.006
Household asset index	-0.005
Father education level	-0.003

Notes: The table reports the correlation of each covariate with the proxy predictor $S(Z)$.

Figure C.1: Corporate Taxes on Entrepreneurship: Lasso Coefficients of Interaction Terms

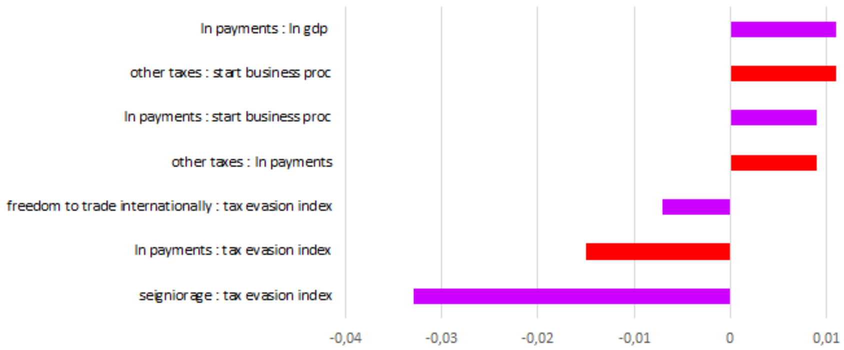


Figure 2.1: $\hat{m}(\cdot)$

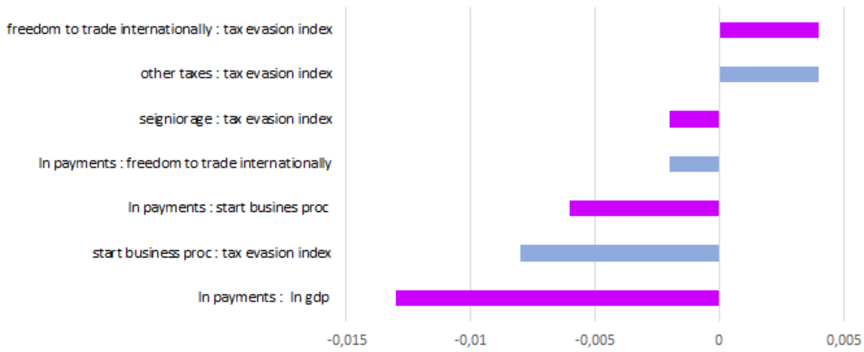


Figure 2.2: $\hat{g}(\cdot)$

Notes: The figure plots the seven largest lasso coefficients of the interaction terms, obtained estimating the nuisance functions $m(\cdot)$ and $g(\cdot)$. Colons indicate interactions of variables. The treatment variable D, is the first year effective corporate tax rate. The dependent variable Y is the average entry rate. The lasso coefficients are calculated as the median over splits.

D Monte Carlo Simulations

To support the general validity of our empirical findings, we perform several Monte Carlo simulations. We evaluate the finite sample performance of the DML technique relative to OLS – the most used estimator in applied economics. Other notable Monte Carlo simulation studies which analyze the finite sample properties of causal ML methods include: Carvalho et al. (2019), Dorie et al. (2019), Hahn et al. (2019), Jacob (2021), Knaus et al. (2018), Wendling et al. (2018). These studies use and compare a variety of causal ML methods under different data generating processes (DGPs). In this study, we depart from their analysis by tailoring our simulations to the data sets we are revisiting and comparing OLS with DML. Thus, we focus on empirically relevant settings in applied economics where the functional forms of the outcome and treatment equations can take both linear and realistic nonlinear forms. In addition, we reproduce typical scenarios concerning the relative sizes of the number of covariates used for estimation, the number of covariates in the DGP and the sample size.

To this end, we consider a framework where the number of covariates included in the estimation by the applied researcher is larger than the number of relevant covariates in the outcome and treatment equations. For instance, this setting is suitable in cases when the applied researcher does not know a priori which are the relevant covariates. We consider a general case in which only some of the relevant covariates appear in the data generating process of both the treatment and the outcome equation.

We start with the linear case, where the true functional forms of both the treatment and outcome equations are linear. In all our linear simulation settings, the true functional form includes ten covariates for the outcome equation $g_0(X)$ and seven covariates for the treatment equation $m_0(X)$. Out of the seven covariates in the treatment equation, four are also present in the outcome equation. We consider a sample size of $n = 100$ with a varying number of covariates included in the estimation, i.e., $k = 20, 50, 90$. Considering a large number of covariates for estimation relative to the sample size is relevant, for instance, when the researcher includes many fixed effects or technical controls. The covariates are drawn from a multivariate normal distribution with means zero, variances one and covariances $0.2^{|r-c|}$, where r is the row index of the variance-covariance matrix of the covariates and c is the column index of this matrix. The error terms U and V are drawn from a bivariate normal with means zero, variances one and covariances zero. The true value of the parameter of interest θ is set to 2. The coefficients of the other covariates randomly take values from the set {-5, -3, -1, 1, 3, 5}. We consider 2 folds and 100 splits for the DML method. Within the DML method, we analyze the performance of the Lasso, Regression Trees, Boosting, Forest, Neural Network and Elastic Net. The values of tuning parameters are the same as in the first two DML applications. The number of simulation repetitions is set to 100.

The nonlinear case has the same setup as the linear case with the exception of the true functional forms of the outcome and treatment models. For these, we have:

$$
\begin{aligned}
g_0(X) &= \sum_{i=1}^{4} c_{1,i} \exp(X_i) + c_{1,5} X_5 \times X_6 + c_{1,6} X_7 \times X_8 + c_{1,7} X_9^2 + c_{1,8} X_{10}^2 \\
&\quad + c_{1,9} \ln(|X_{11} + 1|) \times \ln(|X_{12} + 1|) + c_{1,10} \ln(|X_4 + 1|) + c_{1,11} \frac{1}{X_5} \\
m_0(X) &= c_{2,1} \exp(X_1) \times \exp(X_3) + c_{2,2} X_{13} \times X_{14} + c_{2,3} X_{15} \times X_{16} + c_{2,4} X_{11}^2 \\
&\quad + c_{2,5} \ln(|X_9 + 1|) \times \ln(|X_{17} + 1|) + c_{2,7} \ln(|X_3 + 1|) + c_{2,6} \frac{1}{X_{16}}.
\end{aligned}
$$

The outcome equation $g_0(X)$ contains now four exponential terms, two interaction terms, two square terms, one logarithm term, one interaction of logarithms term and an inverse term. The treatment equations $m_0(X)$ contains an interaction of exponential terms, two interaction terms, a square term, a logarithm term, an interaction of logarithms term and an inverse term. Note that four covariates enter both $g_0(X)$ and $m_0(X)$, but with different functional forms. The type of nonlinearities that we consider are plausible for applied economic studies.

We compute the following evaluation criteria: mean, mean absolute error (MAE), variance (Var), mean squared error (MSE) and OLS hit rate (the proportion of times the OLS estimates are closer to the true estimates compared to the DML estimates). All evaluation measures are computed over the 100 simulation repetitions.

Table D.17 summarizes our results on the linear case, with $n = 100$. Notice that even when the data generating process is completely linear, as the number of covariates used for estimation increases relative to the sample size, the OLS hit

Table D.17: Monte Carlo Simulations: linear case, normal distribution, n=100

	Mean	MAE	Var	MSE	OLS hit rate
			k=20		
OLS	2.030347	0.073340	0.011994	0.012795	–
DML (Lasso)	2.204096	0.308212	0.271943	0.310879	0.800000
DML (Reg. Tree)	1.689803	0.308441	0.121442	0.216449	0.850000
DML (Boosting)	1.980974	0.123810	0.035727	0.035732	0.630000
DML (Forest)	2.017441	0.195251	0.063020	0.062694	0.820000
DML (Neural Net.)	2.179827	0.295220	0.149802	0.180642	0.810000
DML (Elastic Net)	2.023434	0.190474	0.065182	0.065079	0.850000
			k=50		
OLS	1.998316	0.096766	0.020967	0.020760	–
DML (Lasso)	2.305637	0.361159	0.230212	0.321324	0.840000
DML (Reg. Tree)	1.599136	0.465340	0.120046	0.279538	0.870000
DML (Boosting)	1.964103	0.161710	0.076277	0.076803	0.710000
DML (Forest)	2.001251	0.194066	0.064021	0.063382	0.710000
DML (Neural Net.)	2.161565	0.277444	0.120533	0.145431	0.800000
DML (Elastic Net)	1.995786	0.208867	0.068792	0.068122	0.720000
			k=90		
OLS	2.015636	0.232447	0.126762	0.125739	–
DML (Lasso)	2.152435	0.248923	0.150110	0.171845	0.520000
DML (Reg. Tree)	1.644803	0.358840	0.101281	0.226434	0.610000
DML (Boosting)	1.979060	0.172211	0.057695	0.057556	0.430000
DML (Forest)	1.980332	0.181217	0.057536	0.057348	0.400000
DML (Neural Net.)	2.086499	0.238486	0.099034	0.105526	0.500000
DML (Elastic Net)	1.974706	0.169177	0.054517	0.054611	0.400000

Notes: This table reports the results of Monte Carlo simulations where the nuisance functions are linear.

rates decrease. For a sample size of $n = 100$, at $k = 90$, the OLS hit rates start to become smaller than 50%. Thus, if the number of covariates used in estimation is large relative to the sample size, the DML method tends to outperform OLS.

Table D.18 displays the results on the nonlinear case when $n = 100$. Notice that, already at $k = 20$, the OLS hit rates are around 50% or lower, while at $k = 50$, they are below 50% for most of the ML methods. In addition, when comparing the nonlinear and the linear cases with the same number of covariates k, we notice that the OLS hit rates are lower in the nonlinear case.

Furthermore, we perform simulations considering the following cases: 1) larger sample size (n=600); 2) covariates and error terms drawn from uniform distributions: $U[0, 1]^k$ and $U[-1, 1]^2$, respectively. The results are shown in Tables D.19 to D.24, and they further support our main findings for both the linear and the

Table D.18: Monte Carlo Simulations: nonlinear case, normal distribution, n=100

	Mean	MAE	Var	MSE	OLS hit rate
			k=20		
OLS	1.987438	0.074142	0.018767	0.018737	–
DML (Lasso)	1.973458	0.072649	0.024766	0.025223	0.560000
DML (Reg. Tree)	1.933508	0.086160	0.030411	0.034529	0.570000
DML (Boosting)	1.980180	0.084764	0.023779	0.023934	0.540000
DML (Forest)	1.977051	0.073242	0.018031	0.018378	0.450000
DML (Neural Net.)	1.976888	0.067966	0.022784	0.023090	0.540000
DML (Elastic Net)	1.971941	0.061446	0.017436	0.018049	0.520000
			k=50		
OLS	2.036816	0.106510	0.027996	0.029072	–
DML (Lasso)	2.006489	0.082303	0.032754	0.032469	0.410000
DML (Reg. Tree)	1.953161	0.092469	0.041107	0.042890	0.440000
DML (Boosting)	2.070185	0.121154	0.035497	0.040068	0.540000
DML (Forest)	2.017207	0.077788	0.028629	0.028639	0.370000
DML (Neural Net.)	2.010202	0.081012	0.033766	0.033533	0.430000
DML (Elastic Net)	2.004967	0.081342	0.029084	0.028818	0.370000
			k=90		
OLS	1.962616	0.266024	0.167484	0.167207	–
DML (Lasso)	2.008953	0.090226	0.027975	0.027775	0.150000
DML (Reg. Tree)	1.960935	0.102387	0.034753	0.035931	0.230000
DML (Boosting)	2.044449	0.117992	0.037492	0.039092	0.240000
DML (Forest)	2.005692	0.079685	0.025495	0.025273	0.160000
DML (Neural Net.)	2.011866	0.080616	0.027288	0.027156	0.140000
DML (Elastic Net)	1.996920	0.075328	0.024669	0.024431	0.190000

Notes: This table reports the results of Monte Carlo simulations where the nuisance functions are nonlinear.

nonlinear case.

Overall, our results indicate that, if the true data generating process presents nonlinearities similar to our simulation setup, the applied researcher would benefit from using the DML method over OLS. Moreover, the gains of using DML over OLS are larger as the number of covariates used for estimation gets larger. In addition, even when the true data generating process is linear, if the number of covariates used for estimation is large relative to the sample size, the applied researcher would again benefit from using the DML method over OLS.

Table D.19: Monte Carlo Simulations: linear case, normal distribution, n=600

	Mean	MAE	Var	MSE	OLS hit rate
			k=120		
OLS	2.001488	0.026049	0.001833	0.001817	–
DML (Lasso)	1.985497	0.053117	0.010009	0.010119	0.770000
DML (Reg. Tree)	1.607744	0.402505	0.051009	0.204364	0.970000
DML (Boosting)	1.963226	0.149899	0.051153	0.051994	0.890000
DML (Forest)	2.063754	0.164263	0.041947	0.045592	0.880000
DML (Neural Net.)	1.882310	0.116620	0.002952	0.016774	0.930000
DML (Elastic Net)	2.007438	0.073331	0.009803	0.009760	0.760000
			k=300		
OLS	1.999769	0.034485	0.002887	0.002858	–
DML (Lasso)	2.008128	0.099424	0.017887	0.017775	0.830000
DML (Reg. Tree)	1.532333	0.446685	0.035774	0.254129	1.000000
DML (Boosting)	1.964721	0.179759	0.056825	0.057501	0.880000
DML (Forest)	1.973244	0.146126	0.041716	0.042015	0.810000
DML (Neural Net.)	1.914770	0.092823	0.008958	0.016132	0.740000
DML (Elastic Net)	2.084498	0.102074	0.016634	0.023607	0.810000
			k=540		
OLS	1.994620	0.065545	0.016884	0.016744	–
DML (Lasso)	2.069049	0.140462	0.021177	0.025733	0.710000
DML (Reg. Tree)	1.567488	0.418906	0.032098	0.218843	0.920000
DML (Boosting)	1.944381	0.170551	0.055909	0.058443	0.690000
DML (Forest)	2.012795	0.127892	0.034991	0.034805	0.640000
DML (Neural Net.)	1.968579	0.101964	0.024260	0.025005	0.630000
DML (Elastic Net)	2.097952	0.145330	0.022176	0.031549	0.630000

Notes: This table reports the results of Monte Carlo simulations where the nuisance functions are linear.

Table D.20: Monte Carlo Simulations: nonlinear case, normal distribution, n=600

	Mean	MAE	Var	MSE	OLS hit rate
			k=120		
OLS	1.997737	0.032447	0.002612	0.002591	–
DML (Lasso)	1.997321	0.035321	0.002551	0.002532	0.510000
DML (Reg. Tree)	1.973572	0.041298	0.004078	0.004735	0.610000
DML (Boosting)	2.001592	0.035306	0.003748	0.003713	0.440000
DML (Forest)	1.992231	0.031003	0.002387	0.002424	0.530000
DML (Neural Net.)	1.867901	0.133207	0.003541	0.020956	0.900000
DML (Elastic Net)	2.008863	0.034637	0.004966	0.004994	0.520000
			k=300		
OLS	1.992906	0.040230	0.004147	0.004155	–
DML (Lasso)	1.996719	0.031788	0.002323	0.002311	0.410000
DML (Reg. Tree)	1.983927	0.033633	0.002433	0.002667	0.400000
DML (Boosting)	1.993623	0.036100	0.003542	0.003547	0.380000
DML (Forest)	1.991955	0.030786	0.002382	0.002423	0.380000
DML (Neural Net.)	1.965584	0.056685	0.004373	0.005514	0.620000
DML (Elastic Net)	2.001159	0.038427	0.004376	0.004333	0.410000
			k=540		
OLS	2.013503	0.096330	0.016465	0.016482	–
DML (Lasso)	2.007452	0.034834	0.002288	0.002321	0.160000
DML (Reg. Tree)	1.993055	0.036196	0.002699	0.002721	0.200000
DML (Boosting)	1.994680	0.028618	0.002279	0.002284	0.170000
DML (Forest)	2.001690	0.028353	0.002025	0.002008	0.150000
DML (Neural Net.)	2.031335	0.049988	0.004227	0.005166	0.250000
DML (Elastic Net)	2.003342	0.030316	0.002421	0.002408	0.180000

Notes: This table reports the results of Monte Carlo simulations where the nuisance functions are nonlinear.

Table D.21: Monte Carlo Simulations: linear case, uniform distribution, n=100

	Mean	MAE	Var	MSE	OLS hit rate
			k=20		
OLS	1.996376	0.077741	0.012375	0.012265	–
DML (Lasso)	1.901389	0.204322	0.059377	0.068508	0.800000
DML (Reg. Tree)	1.543310	0.462465	0.065671	0.273579	0.910000
DML (Boosting)	2.025640	0.186998	0.068178	0.068154	0.780000
DML (Forest)	2.028667	0.199064	0.064834	0.065008	0.760000
DML (Neural Net.)	1.881087	0.169411	0.043131	0.056839	0.690000
DML (Elastic Net)	1.891030	0.175879	0.042529	0.053978	0.790000
			k=50		
OLS	1.991386	0.118182	0.019108	0.018992	–
DML (Lasso)	1.946033	0.170420	0.082413	0.084502	0.680000
DML (Reg. Tree)	1.552121	0.437326	0.078627	0.278437	0.890000
DML (Boosting)	1.991912	0.191136	0.071496	0.070847	0.590000
DML (Forest)	1.991160	0.185206	0.069593	0.068975	0.600000
DML (Neural Net.)	1.911374	0.169080	0.063565	0.070784	0.650000
DML (Elastic Net)	1.975024	0.160858	0.068118	0.068060	0.640000
			k=90		
OLS	1.994319	0.217047	0.104711	0.103696	–
DML (Lasso)	2.028445	0.233100	0.129348	0.128863	0.540000
DML (Reg. Tree)	1.556620	0.467501	0.074451	0.270292	0.790000
DML (Boosting)	1.981064	0.191275	0.073147	0.072774	0.420000
DML (Forest)	1.984326	0.173800	0.070628	0.070167	0.440000
DML (Neural Net.)	1.949135	0.160300	0.063828	0.065777	0.430000
DML (Elastic Net)	2.064030	0.226524	0.107091	0.110119	0.510000

Notes: This table reports the results of Monte Carlo simulations where the nuisance functions are linear. The covariates and the error terms are drawn from a uniform distribution.

Table D.22: Monte Carlo Simulations: nonlinear case, uniform distribution, n=100

	Mean	MAE	Var	MSE	OLS hit rate
			k=20		
OLS	2.002738	0.064782	0.009416	0.009330	–
DML (Lasso)	2.012751	0.042929	0.007036	0.007128	0.330000
DML (Reg. Tree)	1.958865	0.054497	0.008699	0.010304	0.480000
DML (Boosting)	2.044673	0.051413	0.007330	0.009252	0.420000
DML (Forest)	2.008008	0.043494	0.005324	0.005334	0.360000
DML (Neural Net.)	2.007946	0.067527	0.012634	0.012571	0.600000
DML (Elastic Net)	2.012733	0.045895	0.007138	0.007229	0.330000
			k=50		
OLS	1.975906	0.081323	0.029820	0.030102	–
DML (Lasso)	1.989841	0.040405	0.018834	0.018749	0.320000
DML (Reg. Tree)	1.937210	0.061130	0.018722	0.022477	0.460000
DML (Boosting)	2.023959	0.048177	0.018612	0.019000	0.280000
DML (Forest)	1.997268	0.033068	0.015187	0.015043	0.270000
DML (Neural Net.)	2.052180	0.100297	0.030324	0.032743	0.500000
DML (Elastic Net)	1.990408	0.041291	0.018850	0.018754	0.310000
			k=90		
OLS	2.011864	0.223696	0.127822	0.126684	–
DML (Lasso)	2.002382	0.045577	0.004474	0.004435	0.150000
DML (Reg. Tree)	1.943621	0.053735	0.006174	0.009291	0.180000
DML (Boosting)	2.024395	0.046159	0.003646	0.004205	0.160000
DML (Forest)	1.998078	0.043647	0.003124	0.003097	0.150000
DML (Neural Net.)	2.091630	0.102777	0.015438	0.023680	0.330000
DML (Elastic Net)	2.001142	0.044925	0.004401	0.004359	0.150000

Notes: This table reports the results of Monte Carlo simulations where the nuisance functions are linear. The covariates and the error terms are drawn from a uniform distribution.

Table D.23: Monte Carlo Simulations: linear case, uniform distribution, n=600

	Mean	MAE	Var	MSE	OLS hit rate
			k=120		
OLS	2.004793	0.026902	0.002012	0.002015	–
DML (Lasso)	1.936035	0.081600	0.011072	0.015053	0.760000
DML (Reg. Tree)	1.425837	0.572129	0.051434	0.380583	0.990000
DML (Boosting)	1.919976	0.140301	0.051898	0.057783	0.850000
DML (Forest)	1.938514	0.142365	0.056923	0.060134	0.840000
DML (Neural Net.)	1.779801	0.215857	0.008176	0.056582	0.970000
DML (Elastic Net)	1.876710	0.118719	0.009658	0.024762	0.870000
			k=300		
OLS	2.002833	0.036046	0.002761	0.002742	–
DML (Lasso)	1.948221	0.086708	0.014652	0.017187	0.770000
DML (Reg. Tree)	1.492953	0.517448	0.060832	0.317321	0.970000
DML (Boosting)	1.983007	0.161867	0.054660	0.054402	0.870000
DML (Forest)	2.013929	0.156971	0.056243	0.055875	0.860000
DML (Neural Net.)	1.848317	0.152745	0.021876	0.044665	0.880000
DML (Elastic Net)	1.867112	0.127011	0.014486	0.032001	0.840000
			k=540		
OLS	2.000100	0.061791	0.013361	0.013227	–
DML (Lasso)	1.952425	0.098515	0.015769	0.017875	0.580000
DML (Reg. Tree)	1.486512	0.500889	0.066259	0.329266	0.950000
DML (Boosting)	1.988351	0.169045	0.064906	0.064393	0.700000
DML (Forest)	2.023011	0.193296	0.066738	0.066600	0.730000
DML (Neural Net.)	1.953577	0.153075	0.053910	0.055526	0.700000
DML (Elastic Net)	1.859506	0.132078	0.014241	0.033837	0.660000

Notes: This table reports the results of Monte Carlo simulations where the nuisance functions are linear. The covariates and the error terms are drawn from a uniform distribution.

Table D.24: Monte Carlo Simulations: nonlinear case, uniform distribution, n=600

	Mean	MAE	Var	MSE	OLS hit rate
			k=120		
OLS	1.997828	0.028075	0.004301	0.004263	–
DML (Lasso)	1.998323	0.024324	0.003442	0.003411	0.360000
DML (Reg. Tree)	1.951593	0.058686	0.003763	0.006068	0.680000
DML (Boosting)	2.010452	0.027017	0.001463	0.001558	0.460000
DML (Forest)	1.994384	0.025305	0.001373	0.001391	0.450000
DML (Neural Net.)	1.959658	0.037201	0.004208	0.005794	0.580000
DML (Elastic Net)	1.998315	0.024289	0.003440	0.003409	0.360000
			k=300		
OLS	2.000135	0.031734	0.002869	0.002841	–
DML (Lasso)	2.002880	0.020608	0.001402	0.001396	0.320000
DML (Reg. Tree)	1.960344	0.040298	0.002818	0.004363	0.570000
DML (Boosting)	2.013558	0.022721	0.001104	0.001277	0.370000
DML (Forest)	1.995551	0.021004	0.001145	0.001153	0.340000
DML (Neural Net.)	2.084299	0.079557	0.004390	0.011452	0.740000
DML (Elastic Net)	2.002889	0.020027	0.001399	0.001394	0.320000
			k=540		
OLS	2.021303	0.093354	0.021213	0.021454	–
DML (Lasso)	2.001255	0.023741	0.001074	0.001065	0.160000
DML (Reg. Tree)	1.956112	0.044285	0.003023	0.004919	0.350000
DML (Boosting)	2.017319	0.026946	0.001407	0.001693	0.190000
DML (Forest)	1.997612	0.025841	0.001034	0.001029	0.170000
DML (Neural Net.)	2.123899	0.128300	0.010289	0.025538	0.540000
DML (Elastic Net)	2.001159	0.023787	0.001074	0.001065	0.160000

Notes: This table reports the results of Monte Carlo simulations where the nuisance functions are nonlinear. The covariates and the error terms are drawn from a uniform distribution.