GSC Advanced Research and Reviews

eISSN: 2582-4597 CODEN (USA): GARRC2

Cross Ref DOI: 10.30574/gscarr

Journal homepage: https://gsconlinepress.com/journals/gscarr/

(RESEARCH ARTICLE)

Check for updates

Econometric advances in causal inference: The machine learning revolution

Shuvo Kumar Mallik [1,*], Imran Uddin [2], Sadia Maliha Trisha [3], Md. Morshedul Hasan [4] and M Abeedur Rahman [5]

[1] *Department of Economics, Southeast University, Dhaka, Bangladesh.*
[2] *A2Z Finance Australia (Easy Mortgage Solutions Australia), Australia.*
[3] *Dublin Business School, Dublin, Ireland.*
[4] *School of Business Roya University, Dhaka, Bangladesh.*
[5] *Assistant Professor, Department of Economics, Southeast university, Dhaka, Bangladesh.*

GSC Advanced Research and Reviews, 2025, 22(03), 229-244

Publication history: Received on 08 February 2025; revised on 15 March 2025; accepted on 17 March 2025

Article DOI: https://doi.org/10.30574/gscarr.2025.22.3.0082

Abstract

This is one of the challenges that new and fast-growing econometric literature is beginning to tackle in addressing causal inference problems with machine learning methods. Yet, empirical economics still has not really made use of the strengths of these modern approaches. Here, we revisit groundbreaking empirical work through the perspective of causal machine learning methods to connect econometric theory with applied economics. In particular, we will cover double machine learning, causal forests, and more general machine learning methodologies, both in the setting of average treatment effects and heterogeneous treatment effects. We demonstrate the application of these methods in diverse settings and discuss their significance and additional benefits relative to classical approaches that were utilized in the original studies.

Keywords: Econometric Literature; Machine Learning Methods; Value; Applied economics; Empirical work

1. Introduction

Estimating the causal effect of some intervention on an outcome of interest is one of the main goals of empirical research in economics. However, omitted variable bias can lead to misleading estimates in observational studies, so it is critical to include a wide array of control variables. Even with a relatively small number of raw covariates, interactions and transformations can be added that dramatically increase the number of controls in a regression. In such settings, machine learning (ML) methods are instrumental. However, standard ML prediction models are, at a fundamental level, not designed to achieve the same goals as most empirical economic research. Since ML techniques are optimized to predict outcomes for test samples, model selection is performed so that prediction errors in hold-out test sets are minimized. In contrast, most empirical economic research aims to learn about the underlying relationships that underlie policy decisions, not to improve predictive accuracy. This fundamental distinction is why you cannot unthinkingly use techniques from the ML world, which is fundamentally about predicting, to causal inference problems, a misuse that can yield biased estimates.

Still, converging econometric literature is making impressive progress in applying ML methods to causal inference (Athey et al., 2018; Chernozhukov, Chetverikov, Facchinei et al., 2018; Chernozhukov, Demirer, et al., 2018; Wager & Athey, 2018). This literature offers new insights and theoretical results that speak to both ML and econometrics/statistics. Yet, so far, modern causal estimation techniques have not been fully adopted by empirical economics. This paper aims to showcase researchers with empirical evidence on the benefits of causal machine learning context in the real world. We do this by revisiting a number of influential studies applying causal ML methods, and we

* Corresponding author: Shuvo Kumar Mallik

compare the resulting estimates to those produced using conventional econometric approaches. We analyze average treatment effects (ATE) and heterogeneous treatment effects (HTE).

Our primary contribution is to demonstrate the broad applicability of causal ML methods and show the additional value and relevance of these methods from a standard econometric perspective. To validate our results, we also run a number of Monte Carlo simulations in which the accurate data-generating process is known. This enables us to contrast the small-sample performance of various causal ML estimators against that of their traditional econometric estimator counterparts under conditions analogous to those of the studies under reevaluation. For ATE estimation, we utilize the double/debiased machine learning (DML) method introduced by Chernozhukov et al. (2017). For HTE, we use Athey et al.'s causal forest method. 2018) and Wager & Athey (2018) , and the generic ML approach to heterogeneous treatment effects was developed by Chernozhukov, Demirer et al. (2018). These are state-of-the-art causal ML methods with a solid theoretical foundation.

We revisit a series of seminal studies on key applied economics topics published in top journals like the Quarterly Journal of Economics and the American Economic Journal: Macroeconomics and the American Economic Journal: Applied Economics in the relatively recent past. We limit our analysis to studies that provide the entire dataset used for replication purposes in a publicly accessible format, either on the journal's website or the authors' web pages. We look back at two observational studies for ATE estimation: Jankov et al. (2010a) on corporate taxes and investment and Nunn & Trefler (2010a) on skill-biased tariffs and long-run growth. To demonstrate the applicability and efficacy of each method, we consider two different empirical studies relevant for HTE estimation: we extend the analysis of DellaVigna & Kaplan (2007a), which demonstrates how Fox News affects the Republican vote share and analyze the effect of a teacher training intervention on student performance in a randomized experiment analyzed by Loyalka et al. (2019a). Each of these studies includes a rigorous econometric analysis of their primary research question, an analysis we do not aim to replicate fully. Therefore, we focus on readdressing the above key questions through causal ML methodologies.

From the outcomes of our sample of revisited studies, we distilled and organized four theoretical rationales outlining how causal ML methods add value to causal analysis relative to traditional econometric techniques. These benefits are not confined to the specific studies or datasets we consider but are generalizable across empirical economic research.

First, causal ML approaches supply potent methods for discovering complex interactions among variables and for flexibly estimating the relationships between outcomes, treatments, and covariates. This is especially relevant in observational studies as the assumption that treatment is as good as randomly assigned conditional on observables is crucial. If some covariates are correlated with both the treatment variable and the outcome, estimates of treatment effects will be biased when not conditioning on all relevant confounders. For instance, in the paper on the impact of corporate taxes on investment and entrepreneurship, the original analysis by Jankov et al. (2010a) showed a strong negative effect of corporate taxes. However, their results are attenuated when conditioning on all possible controls together.

In contrast, implementing DML obtains more significant effect estimates than the ones presented by Jankov et al. (2010a), which continue to be statistically significant. In a similar vein, our estimates of the impact of skill-biased tariffs on growth imply a minor effect compared to what was initially found by Nunn & Trefler (2010a). In some specifications, the estimates themselves are no longer statistically significant. These findings would imply that DML estimates are more robust to possible nonlinear confounders.

Second, causal ML methods can be helpful when the number of covariates is large relative to the sample size. In contrast to standard econometric methods that treat all covariates as potentially relevant, ML methods make a sparsity assumption where, in fact, only a small set of covariates matter and use regularized regression approaches. In publications like Jankov et al. (2010a) and Nunn & Trefler (2010a)), there are many raw covariates relative to the sample size, and it is infeasible with traditional methods to include all possible nonlinear transformations and interactions. Nunn & Trefler (2010a), for example, restrict themselves to logarithmic transformations; Jankov et al. (2010a) do not have any nonlinear terms. In contrast, DML enables us to flexibly control all relevant confounders expressed in both linear and nonlinear forms at once.

Third, causal ML methods can promote systematic model selection. The most prominent machine learning branch methods involve estimating and comparing thousands of possible model specifications with a search for the best functional forms for the available data. For instance, the corporate tax effects we find are consistent with Jankov et al. (2010a). Still, we demonstrate that DML model selection, the data-driven selection of a small subset of influential covariates from a much larger set of potential controls, results in more significant coefficient estimates and more minor

standard errors than OLS (ordinary least squares) regressions where all the covariates are included. Conventional model selection processes often involve ad hoc decisions around model specification that can yield different point estimates and, therefore, lead to different policy recommendations. We also demonstrate how causal ML methods can serve as robustness checks or supplementary analyses. Standard robustness checks usually consist of the presentation of several alternative regression specifications, whereas ML-based methods make sure that essential covariate transformations are not missed. As an example, our reanalysis of Nunn & Trefler (2010a) is just a robustness check to the extent that DML estimates are flexible enough to suit data-driven functions of covariates. Our findings diverge from the original analysis, and they lose statistical significance in this case.

Finally, causal ML techniques are of great use for estimating heterogeneous treatment effects. Since causal ML methods are able to systemically control for many covariates that can make treatment effects heterogeneous, they are less likely to omit relevant interactions than standard econometric techniques. Our analyses of the impact of Fox News on Republican vote share (following DellaVigna & Kaplan, 2007a) and of teaching training interventions (Loyalka et al., 2019a) provide an example of this advantage: we identify sources of heterogeneity revealed by our results that were not highlighted by the original studies. Additionally, many causal ML methods used for estimating HTE have been shown to produce valid confidence intervals in high-dimensional settings44,45, where conventional p-values resulting from individual hypothesis tests may be invalid due to multiple hypothesis testing issues46, which are commonly faced when investigating treatment effect heterogeneity.

1.1. Average Treatment Effects

Examining Investment and Entrepreneurship: Corporate Taxes (Jankov et al., 2010a) and Skill-Biased Tariff Externalities and Economic Growth (Double Machine Learning (DML)) (Chernozhukov et al., 2018).

1.1.1. The Effect of corporate taxes on investment and entrepreneurship

Description of original analysis.

The first study applying causal machine learning methods that we examine looks at corporate tax investment and entrepreneurship (Jankov et al., 2010a). This is an observational study based on estimates of OLS country-level regressions for 2004, controlling for a range of changes in corporate tax rates. The study finds that corporate taxes depress investment and entrepreneurship. The sample consists of 50 to 85 countries, depending on specification.

Of these, the original paper considers four outcome variables: investment as a share of GDP, FDI as a share of GDP, business density per 100 people, and new firm average entry rate. The statutory corporate tax rate, adequate first-year corporate income tax liability for a new firm, and a firm-specific tax rate that is consistent with the actual depreciation schedule over five years.

The results from the original study reports are based on multiple regression specifications, which involve different groups of control variables to consider potential confounders related to corporate tax rates. The key variables that drive the results are based on the findings from these regressions. Jankov et al. (2010a) show regression results with the first two sets of covariates added independently. We perform a final robustness check that adds all the control variables (a total of twelve) into the same regression.

For specifications where only one control set is included at a time, the study reports a negative and statistically significant effect of corporate taxes on entrepreneurship and investment. When all control variables are aggregated together, the negative relationship persists, but the coefficients decrease in magnitude and lose statistical significance.

1.2. DML analysis

We return to the last robustness check in the paper, one which uses all four sets of covariates at once in the same DML partially linear model. The findings are summarized in Table 1. Columns (1) through (7) show the DML point estimates of the effect of corporate taxes on investment and entrepreneurship, using a variety of ML methods to estimate the nuisance functions. Importantly, all DML point estimates are pessimistic and remain on a similar scale for all ML approaches.

We then compare the results with the original paper's results, which are displayed in column (8) of the regression table, and present the complete set of covariates; we find that the absolute magnitude of the DML coefficients are typically more prominent, while standard errors are typically more minor in almost all regressions. Moreover, about half (40 out of 84) of the regressions (at least) reach the standard of significance at the 10 level. This indicates that regularization

helps to lower estimation errors and increase reliability. Nevertheless, since no ground truth is available, whether DML estimates are close is an open question.

We have reached this conclusion based on the fact that both the original analysis and our method of adjustment are built on assumptions about confounding factors. OLS regression assumes that all relevant factors are controlled for with a linear adjustment. In comparison, DML provides greater flexibility in that both linear and nonlinear controls can be applied for possible confounders. In other words, DML loosens the substantial restrictions of the original analysis in that it only assumes a weaker assumption that confounders can be controlled to a sufficient degree by adding both linear and nonlinear terms using the same set of controls.

We are also interested in a posteriori finding the exact nonlinear terms that cause differences in the estimates. This may be a challenging task because ML methods like neural networks (and hybrid methods) are used to estimate nuisance functions. One way could be to see the LASSO non-zero coefficients and try to see if they capture any nonlinear effect.

The DML results come from cross-validating ML parametrization whenever it is mathematically sensible to do so. But, not all parameters are data-driven, such as the number of trees or leaves in a single tree. We also perform additional robustness checks for these non-adaptive tuning parameters. Moreover, we also change the activation function and the number of layers of neural networks. Results are available upon request and do not differ from those presented here.

The assumption of sparsity makes causal machine-learning methods effective. However, this assumption cannot be tested directly and should be used with care. This finding of highly similar second-stage DML estimates, regardless of ML methods, supports our hypothesis that a stable underlying pattern can be identified and is consistently captured across ML techniques for our empirical application (ATE estimates).

This application of causal ML to the real world is a good example that depicts the trade-offs we make with causal ML. On the one hand, researchers try to control for as many confounders as they can so as to make their assumptions more credible. On the contrary, when controlling for a long list of covariates, especially in small samples, overfitting will take place, and standard errors will be inflated. In this case, the original authors used a "kitchen sink" regression where they put all covariates in the same model at once, giving them more significant standard errors than what we get. The DML method assists in managing this trade-off as it strengthens the credibility of confounding assumptions (by flexibly modeling confounder effects) and implements a data-driven variable selection strategy. This method reduces the number of potential controls to only those confounders that are influential, leading to less estimation error and improved precision.

Table 1 The effect of corporate taxes on investment and entrepreneurship

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
	Lasso	**Reg. Tree**	**Boosting**	**Forest**	**Neural Net**	**Ensemble**	**Best**	**OLS**
Panel A: Investment 2003–2005								
Statutory corporate tax rate	−0.081	−0.056	−0.065	−0.077	−0.056	−0.074	−0.068	−0.064
	(0.083)	(0.075)	(0.076)	(0.084)	(0.103)	(0.09)	(0.089)	(0.098)
First-year effective tax rate	−0.122	−0.133	−0.156	−0.142	−0.137	−0.134	−0.138	−0.117
	(0.092)	(0.089)	(0.087)	(0.093)	(0.101)	(0.091)	(0.091)	(0.106)
Five-year effective tax rate	−0.178	−0.179	−0.199	−0.204	−0.218	−0.195	−0.203	−0.189
	(0.096)	(0.095)	(0.091)	(0.094)	(0.101)	(0.099)	(0.101)	(0.118)
Observations	61	61	61	61	61	61	61	61
Panel B: FDI 2003–2005								

Statutory corporate tax rate	− 0.136	− 0.167	− 0.142	− 0.131	− 0.078	− 0.123	− 0.112	− 0.030
	(0.085)	(0.088)	(0.09)	(0.091)	(0.09)	(0.092)	(0.092)	(0.066)
First-year effective tax rate	− 0.172	− 0.203	− 0.188	− 0.169	− 0.154	− 0.168	− 0.16	− 0.1
	(0.091)	(0.084)	(0.085)	(0.07)	(0.084)	(0.088)	(0.085)	(0.071)
Five-year effective tax rate	− 0.162	− 0.183	− 0.169	− 0.177	− 0.164	− 0.17	− 0.15	− 0.095
	(0.093)	(0.076)	(0.076)	(0.08)	(0.09)	(0.086)	(0.084)	(0.081)
Observations	61	61	61	61	61	61	61	61
Panel C: Business density								
Statutory corporate tax rate	− 0.054	− 0.088	− 0.063	− 0.06	− 0.031	− 0.054	− 0.042	− 0.034
	(0.063)	(0.072)	(0.066)	(0.063)	(0.077)	(0.067)	(0.069)	(0.083)
First-year effective tax rate	− 0.105	− 0.158	− 0.123	− 0.115	− 0.091	− 0.099	− 0.102	− 0.068
	(0.074)	(0.087)	(0.073)	(0.07)	(0.083)	(0.074)	(0.076)	(0.092)
Five-year effective tax rate	− 0.093	− 0.14	− 0.11	− 0.104	− 0.087	− 0.107	− 0.098	− 0.070
	(0.075)	(0.085)	(0.072)	(0.068)	(0.086)	(0.076)	(0.075)	(0.103)
Observations	60	60	60	60	60	60	60	60
Panel D: Average entry rate 2000–2004								
Statutory corporate tax rate	− 0.128	− 0.15	− 0.141	− 0.133	− 0.079	− 0.12	− 0.113	− 0.029
	(0.067)	(0.066)	(0.066)	(0.065)	(0.081)	(0.071)	(0.071)	(0.086)
First-year effective tax rate	− 0.107	− 0.136	− 0.14	− 0.115	− 0.109	− 0.116	− 0.112	− 0.083
	(0.075)	(0.066)	(0.069)	(0.066)	(0.082)	(0.074)	(0.072)	(0.094)
Five-year effective tax rate	− 0.156	− 0.146	− 0.155	− 0.15	− 0.175	− 0.155	− 0.152	− 0.133
	(0.076)	(0.072)	(0.072)	(0.07)	(0.087)	(0.075)	(0.077)	(0.103)
Observations	50	50	50	50	50	50	50	50
Raw covariates	12	12	12	12	12	12	12	12

Notes: 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.

2. The effect of skill-biased tariffs on growth

2.1. Description of original analysis.

Nunn and Trefler (2010) investigate the impact of such skill-biased tariffs, which disproportionately favor skill-intensive industries, on the long-run growth of economies. Developing the theoretical model of Grossman and Helpman (1991)2, the authors show that it is possible to have tariffs that benefit skill-intensive industries, then expansion in these industries, and higher sustainable growth.

The paper uses both cross-country and industry-level data to provide some empirical evidence of the positive relationship between skill-biased tariffs and long-run growth, as well as investigate channels through which this relationship works. While the theoretical framework explains some of the overall correlation found between skill-biased tariffs and output growth, it is not able to explain the relationship entirely. The paper further explores the persistence of skill-biased tariffs and the importance of institutional quality. In particular, it cites the presence of more robust institutions within a country as being more likely to enable the implementation of policies to protect skill-intensive industries.

Nunn and Trefler (2010) use three measures of skill bias in tariffs in the earlier period in their analysis: (1) the correlation of tariffs with the skill intensity of 12 industries, (2) the log of average tariffs in skill-intensive industries, and (3) the difference in log average tariffs between skill-intensive industries and less skill-intensive industries.

The analysis is conducted at the country level to assess the effect of these tariff measures on annual per capita GDP growth, controlling for a panel of covariates in our regression models. The dependent variable is the average annual log change in industry output in a country for industry-level estimates, and the regressions control for the same set of control variables as the country-level models but also for industry-fixed effects. It contains data on 1,004 observations in 59 countries.

In addition to these, some specifications of the model contain further variable initial tariffs in the industry of interest to account for a potential transmission mechanism. An intuitive idea behind this is that skill-biased tariffs will reallocate resources into skill-stretching sectors, which enjoy positive spillovers, generating higher rates of sustainable economic growth in the long run. If this mechanism is also in effect, industries subject to higher initial tariffs should experience relatively more output growth in the long run. So, suppose this channel is what explains the effect of skill-biased tariffs on growth. In that case, the inclusion of this variable in the regression should reduce the estimated coefficient of tariff skill bias.

2.2. DML analysis.

In summary, Table 2 shows findings from the DML partially linear model approach using country-level data. For all the machine learning methods and the three treatment variables we analyze, we find DML estimates of treatment effects that are much smaller in magnitude than those derived in the original paper. Moreover, the majority of the estimated effects are not statistically significant, with the exception of the coefficients estimated with LASSO. Panel B: boosting, and Panel C: ensemble methods, 10% level of statistical significance.

Table 2 The structure of tariffs and long-term growth: country-level estimates

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
	Lasso	**Reg. Tree**	**Boosting**	**Forest**	**Neural Net**	**Ensemble**	**Best**	**OLS**
Panel A: Skill tariff correlation								
Skill tariff correlation	0.019	0.016	0.016	0.016	0.013	0.019	0.016	0.035
	(0.010)	(0.012)	(0.011)	(0.011)	(0.015)	(0.012)	(0.011)	(0.010)
Panel B: Tariff differential (low cut-off)								
Tariff differential (low cut-off)	0.010	0.008	0.009	0.008	0.006	0.008	0.008	0.016
	(0.005)	(0.005)	(0.005)	(0.006)	(0.008)	(0.006)	(0.006)	(0.006)
Panel C: Tariff differential (high cut-off)								
Tariff differential (high cut-off)	0.009	0.006	0.007	0.008	0.013	0.009	0.008	0.02
	(0.005)	(0.005)	(0.005)	(0.005)	(0.008)	(0.005)	(0.005)	(0.004)
Observations	63	63	63	63	63	63	63	63
Raw covariates	17	17	17	17	17	17	17	17

Notes: Analysis of 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.

At the industry level, all estimation methods show that the treatment results are statistically insignificant apart from the encouragement estimates. The general point is that the DML results indicate that the positive relation between skill-biased tariffs and long-run economic growth does not hold when controlling for average tariff level, country characteristics, date of the initial production structure, and unobserved heterogeneity driving cohort and region fixed effects. In fact, the trivial size of the DML estimates suggests that OLS regressions are deficient in controlling for unobserved confounding. The original paper should be credited with the paper's use of the correlation between skill-biased tariffs and long-term growth, which is explained so much in terms of the character of such tariffs being higher, where institutions are more substantial. These exact country-level DML estimates confirm the hypothesis in this case and suggest that the direct effect of the tariffs' skill bias is less pronounced than would seem the case according to OLS regressions.

Finally, as our results relate to the relationship between skill-biased tariffs and long-run growth, we are not concerned with the relationship between skill-biased tariffs and institutions or the relationship between institutions and long-run growth, both of which were studied in the original paper. Thus, our results are in agreement with the alternative explanations by Nunn and Trefler (2010), and that a causal relationship is present between institutions and economic growth.

2.2.1. Heterogeneous Treatment Effects

Introducing Heter Hawkes Methods and Applications One shows yet heterogeneous treatment effects (HTE) for studies of two using advanced machine lessons. It examines the effects of Fox News on Republican voting behavior, as analyzed by DellaVigna and Kaplan (2007a), through the lens of a causal forest approach (Wager and Athey, 2018; Athey et al., 2019). It also investigates the impact of a teacher training intervention, as examined by Loyalka et al. (2019a)), in the context of a general machine learning framework (Chernozhukov, Demirer et al., 2018).

2.2.2. The effect of Fox News on the republican vote share

Description of original analysis.

Specifically, we revisit and analyze more closely the study of media bias in affecting voting by DellaVigna and Kaplan (2007a). The paper focuses on answering this question by studying the impact of Fox News (a right-leaning cable television channel) on the United States' Republican Party vote share. To measure the effect of Fox News on electoral outcomes, the authors compared whether cities that gained access to Fox News between 1996 and 2000 showed an increase in Republican vote share during that period. It uses data at the level of the city for 9256 cities. We brief on our primary outcome variable: the relative change in GOP vote share between 1996 and 2000. The treatment variable denotes whether Fox News aired in a city between 1996 and 2000. The regression models include a number of controls to help account for potential confounding factors.

DellaVigna and Kaplan (2007a): The Impact of Fox News on the Republican Vote Share They also examine heterogeneity in this effect along city characteristics, including the number of cable channels available, urban population share, and whether the city was situated in a swing or Republican-dominated district. They do so by including regressions with interaction terms between the treatment variable and these covariates.

3. Causal forest analysis

The analysis of heterogeneity in treatment effects (HTE) is carried out using the causal forest approach. It is essential to investigate heterogeneous treatment effects to ascertain if there are effect modifiers based on some features of cities or districts. Although the average treatment effects reported for Fox News across the sample as a whole offer respectably informative measures, treatment effects are rarely homogeneous across cases. Perhaps the impact of Fox News was limited to a handful of areas. There are several ways in which the processes that produced the most and least vigorous responses might differ, allowing us to discern their respective characteristics. There are two primary goals for this exercise. First, we take a data-agnostic view of heterogeneity and test if the treatment effect is conditioned by city or district characteristics. Second, we check to see if the HTE found in the original paper agrees with what causal ML finds.

We show results from the second of the two preferred specifications in the original paper, which also include district fixed effects. We show results for two versions of causal forests that correctly account for district-level effects in different ways. The first set of results includes dummy variables for the congressional district where each city lies. In the second set of results, we implement a cluster-robust version of the Athey and Wager (2019) random forest, where we take each district as a separate cluster. This cluster-robust approach to causal forests relies on the reasonable assumption that clusters do not have an additive effect on the outcome.

Table 3 Fox News—Causal Forest: average treatment effects and test for heterogeneity.

	(1)	**(2)**
	District dummies	Cluster-robust
Fox News effect (ATE)	0.0065	0.0065
	(0.0016)	(.0026)
Fox News effect above median	0.011	0.0078
	(0.0023)	(0.0028)
Fox News effect below median	— 0.0028	0.0034
	(0.0022)	(0.0042)
95% Confidence interval for the difference	(0.00759, 0.01985)	(– 0.00545, 0.01437)
Observations	9,256	9,256

Notes: This table shows the average treatment effect estimate and a test for overall heterogeneity from the causal forest. Standard errors are in parentheses. asymptotic normality for the causal forest under the assumption that the number of covariates is not too large, and the covariates are continuous. In order to avoid this problem, we conduct a robustness check using Athey and Wager (2019)'s method, wherein we fit a first random forest on all covariates, and we then fit a second random forest on a smaller set of variables.

Begin with average treatment effects. These are shown in Table 3. Just like in the original analysis, we see a positive and statistically significant effect of Fox News on Republican vote share when we include district dummies and apply the cluster-robust causal forest, but in this case, the cluster-robust forest results in fatter a variance/covariance matrix. We find that municipalities with access to Fox News had a 0.65-percentage-point higher share of votes cast for Republican candidates than those in cities where Fox News was not accessible.

We would next like to check if the causal forest can indeed capture the heterogeneity of the treatment effects. In Athey and Wager (2019), it is shown that we can partition observations according to whether their estimated bag-conditioned average treatment effect (CATE) is greater or less than the median and compute the average treatment effect separately for these two groups. These are shown in Table 3 as the Fox News effect for below and above the median. Note that this interpretation has to be taken with caution since the valid standard errors for causal forests are an open question, and we are not using predictions that are incorrect for our subgroup creation. The gap between the two subgroups could, therefore, be due to heterogeneity, and that may be connected with the fact that when the two assumptions do not produce a zero (see column 1 of Table 3). However, the same heuristic test performed with the cluster-robust forest does not reveal significant variation in treatment effects. Such a finding may suggest that the variability of the model with district dummies has been overestimated (since the dummy variables are likely not capturing district-specific effects). The cluster-robust causal forest captures district-specific effects more flexibly and may better suit this scenario.

Table 4 Fox News—causal forest: HTE analysis.

	(1)	**(2)**	**(3)**
	CATE below median	CATE above median	*p*-value difference
Panel A: District dummies			
Employment rate, diff. btw. 2000 and 1990	0.00929	0.0005	0.00562
	(0.00243)	(0.00204)	
Share high school degree 2000	0.00806	— 0.00032	0.00676
	(0.00222)	(0.00216)	
Decile 10 in no. cable channels available	0.00872	— 0.00456	4e-05
	(0.00191)	(0.00262)	
		Panel B: Cluster-robust	

Employment rate, diff. btw. 2000 and 1990	0.00939	0.00013	0.05676
	(0.00258)	(0.00412)	
Share high school degree 2000	0.0085	— 0.0015	0.05492
	(0.00301)	(0.00425)	
Decile 10 in no. cable channels available	0.0086	— 0.00524	0.01823
	(0.00284)	(0.00513)	

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.

Though the test results for variation overall are ambivalent, there may still be differences across specific covariates. So, we were involved in checking whether some of the covariates could be a possible source of this variability. To achieve this, for each of our variables, we break the sample into two groups: one for those whose value of the covariate of interest is above the median, and the other whose covariate value is less than or equal to its median and we estimate the average treatment effect for each of the two subsamples. HTE results for the variables that seem to be necessary determinants of variation in both specifications are presented in Table 4. As shown in Table 4, three variables are significant determinants of variation (at least at the 10% level) in both specifications: changes in employment between 1990 and 2000, the share of the population with a high school degree or equivalent, and the tenth decile of the number of available cable channels. We find that the effect of Fox News on Republican voting was more substantial in cities with low employment growth (1990 to 2000). Areas of slower economic growth (and smaller hiring increases) that had been under a Democratic president (Bill Clinton) leading up to the 2000 presidential vote were more readily swayed to vote Republican. We further see a more substantial Fox News effect in the cities below the median in terms of the share of the population that has graduated high school (or equivalent). We also observe a more significant positive effect in municipalities that fall into the tenth decile of the number of cable channels available below the median compared to those that exceed the median in this variable, for which the effect is negative and small.

Next, we examine if variation findings in the original paper are replicated by causal forest. DellaVigna and Kaplan (2007) establish a more significant effect of Fox News on Republican voting in towns with fewer cable channels, including district-fixed effects. Although we do not find much variation in our results with this variable, our results for the tenth decile of a number of cable channels substantiate the results from the original analysis, which found that in regions of high competition among cable channels, the Fox News effect is attenuated. The number of cable channels is also variable, with our most significant importance scores in either specification suggesting that this variable may be relevant for variation again.

When investigating variation across political orientation of districts, we confirm DellaVigna and Kaplan's (2007) results: we do not find any significantly different effect for swing districts, and we do not find a significantly smaller effect for Republican districts. However, in our co-location analysis, we find no significant difference in the effect of Fox News between rural and urban cities, although this is the only variation result that remains consistent across all specifications in DellaVigna and Kaplan (2007).

Overall, our analysis of HTE for Fox News on Republican voting is consistent with some of the results in DellaVigna and Kaplan (2007), such as the presence of variation with a number of cables and the absence of more solid effects in districts with alternate political leanings. However, unlike the earlier paper, we see no variation across districts. The causal forest analysis highlights additional unexplored heterogeneity, including more significant effects in cities that had slower employment growth and cities with a lower share of the population with a high school diploma. Lastly, the causal forest retains more of the variation on treatment effects as demonstrated compared to the cluster robust version with district dummy variables when we implement the overall variation test and when analyzing HTE for individual covariates. However, the district dummy variable model may overstate variation relative to the cluster-robust forest model if the dummy variables do not sufficiently describe district-specific effects. This emphasizes the importance of being more careful about clustered observations when using causal forests in empirical applications (Athey and Wager, 2019).

4. The effect of teacher training on student performance

4.1. Description of original analysis

We report the results of a large-scale randomized trial examining the impacts of a teacher professional development (PD) program on student achievement and student and teacher outcomes in China. Loyalka et al. initially studied the trial. (2019). The intervention included over 300 math teachers in a province employed in different schools. (1) PD only; (2) PD with an extra follow-up that included additional components and activities for the trainees; (3) PD plus an assessment of what the teachers remembered from the training sessions; or (4) no PD (control group). Lectures and discussions were used for PD intervention.

Randomization occurred at the school level, and one teacher from each school was randomly assigned to receive the intervention. Primary outcomes were extracted from cross-sectional regression estimates with the treatment variable as a dummy for whether a school was assigned to the treatment arm. Data were collected at three-time points: baseline, midline, and endline. Assessments were made at midline or end line, and the primary outcome was student achievement on mathematics tests. Adjusted for student level, teacher level, and class size control variables

The main paper shows that after one academic year, there is no statistically significant impact of the PD intervention on student achievement — neither for the PD-only nor for the PD-follow-up and/or evaluation treatments. The authors further find no impact on additional outcomes, including measures of teacher knowledge or student motivation. Several possible explanations are offered for the lack of program effectiveness: the content was overly theoretical, the professional development (PD) was delivered passively, and teachers experienced barriers to implementing the recommended practices in schools. The paper also examines heterogeneous treatment effects by interacting the treatment variable with different student and teacher attributes—household resources, baseline achievement levels, amount of training teachers had received prior to the intervention, whether teachers had a college degree, and whether teachers specialized in mathematics. This suggests the effect of the intervention on student achievement may vary depending on teacher characteristics, but there are no heterogeneous effects based on student characteristics.

4.2. Generic ML analysis.

In this paper, we complement the heterogeneous treatment effects (HTE) from both the main paper and apply the generic machine learning method of Chernozhukov, Demirer et al. (2018). Because small and trivial estimates for HTE could hide substantial heterogeneity and variance, exploring heterogeneous treatment effects is particularly relevant for this intervention. So, we aim to provide a more in-depth explanation of the analysis of heterogeneous treatment effects. We address two questions: The first is the extent to which treatment effects are heterogeneous; the second is the extent to which causal machine learning methods can bring good insights about who received the program and who didn't through their systematic search for heterogeneity over a large number of covariates, relative to standard approaches employed in the main paper.

In our analysis, we consider student achievement in mathematics as our primary outcome of interest. Due to the fact that the results in the main paper consistently show near-zero differences when contrasting the control group with the three unique treatment arms, we exclusively analyze the treatment arm associated with the PD intervention and evaluation here. Our analysis is based on a sample of 10,006 students across 201 schools. We use the same methodology as Loyalka et al. (2019) and cluster standard errors at the school level. Besides the complete set of controls shown in the main paper, we also add other variables to our analysis that could act as moderators of the treatment effects: baseline values of several student-level variables that assess behaviors in the classroom from the student's perspective.

Table 5 Teacher training—generic method: best linear predictor.

	(1)	**(2)**
	ATE (β_1)	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	10,006	10,006

The generic method can be combined with a variety of ML tools, including Chernozhukov, Demirer et al. (2018). An examination of whether overall variation in treatment effects can be identified in the first of that two-step process used to evaluate the performance of various methods of machine learning, which include the Best Linear Predictor or BLP and Babel List EPredictors. Table 5 shows the results for Ketter's BLP. In accordance with the main paper, the estimated ATT is small (the estimated effect of PD is 0.002 SD) and is not significantly different from zero (β1 is not significant). But we also get a more extensive and significantly nonzero estimated β2, meaning that there is heterogeneity in the treatment effects. We first estimate group average treatment effects (Gates). We split the sample into five buckets based on the S(Z) quintiles of the ML proxy predictor. Both estimates are significant at the 10% level. The gap between the top vs. the bottom quintiles is notable, which confirms the existence of heterogeneity in treatment effects. In addition, Figure 1 presents the gates' estimates with the 90% confidence intervals for the five quintiles as well as for the entire sample (where ATT is shown as a dark dashed line and confidence intervals are light dashed lines). Note that for the three middle quintiles, the impact of such teacher training interventions is not significantly different from zero.

Then comes classification and regression analysis (CNA), which analyzes the potential sources of variation. In particular, we further investigate the lower and upper quintiles of ATT, for which the PD intervention has a negative and positive effect, respectively. In particular, we contrasted the characteristics of students and teachers in the two groups. Table 6 shows that, given the large number of covariates available, we proceed with ten covariates that have the most mutual correlation with the proxy predictor S(Z).

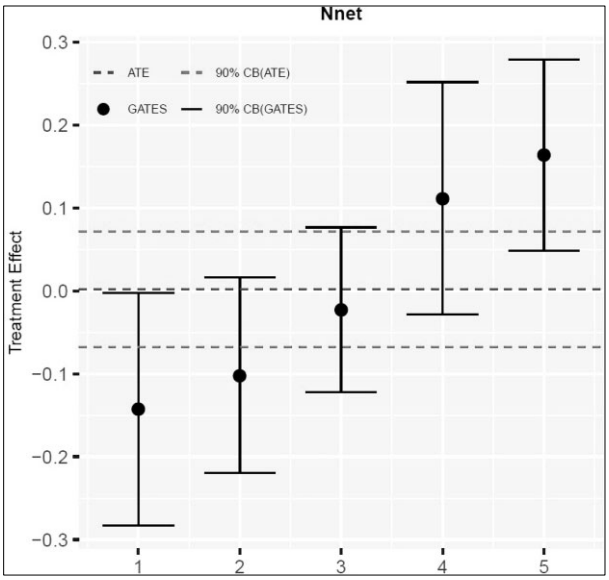

Figure 1 Teacher training—generic method: GATES.

Note: 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.

First, we examine the characteristics of teachers whose students fall in the lowest and highest impacted groups. Notably, the variable most significantly correlated with the proxy predictor was whether the teacher had a college degree; this was also the variable that moderators treatment effects across all treatment arms in the paper. Students in the top 20 percent are more likely than students in the bottom 20 percent to be taught by a teacher with no college degree. Our finding is in agreement with the findings of Lowellka et al. (2019), who found that the intervention had a negative impact on students whose teachers were college-educated, but a positive impact on students whose teachers were less qualified. So that PD might help less qualified teachers, but for more qualified teachers, the costs of intervention (ie, absence from the classroom while attending interventions) outweighed any potential benefits of the intervention itself for the students they were leading (10). The "twilight" of data-driven governance that comes from a degree (cf. the analysis of the generic approach) seems also to drive variation back through the generic approach the PD plus evaluation in the original paper, the one we're all looking at here it seems to matter here a whole lot. The direction of the effect is that students in the top quintile are more likely to be taught by a teacher who lack a major in mathematics, than students in the bottom quintile, and this direction is aligned with the findings from the main analysis.

Table 6 Teacher training—generic method: classification analysis

	(1)	**(2)**	**(3)**
	20% most affected	20% least affected	*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.

Their prior training was received prior to the intervention—although we didn't identify it as influencing variance in the original paper—was more remarkable for the most positively affected group than the least affected group. This may suggest that teachers with more prior training were more successfully able to engage with the PD intervention implementation recommendations. As shown in Table 6, teacher rank, experience, and age in the most positively impacted group are also higher compared to those in the least affected group. This is consistent with the notion that teachers with more experience may be better equipped to implement the strategies taught in the PD intervention effectively. As the PD was primarily theoretical, previous training or more teaching experience could have contributed to their effective transformation of the practices into daily application.

Now, we consider what student characteristics contribute to this heterogeneity. Unlike the results of Lowellka et al. (2019), who did not disentangle variation based on student covariates, we show that students in the most positively affected group differ along multiple dimensions from students in the least affected group. Table 6 includes the factors that have the highest correlation with the heterogeneity score, including student age and gender. The average student in the most positively affected group is approximately half of a year older than the average student in the least affected group, and a higher proportion of these students are male. The most impacted group also has lower baseline math ability and more significant anxiety about mathematics. The findings indicate that PD could benefit students with low prior knowledge and high anxiety in the subject area.

Class size also seems to be a potential determinant of variation. The students who benefit more from PD tend to be in smaller classes. These findings imply that, in a smaller classroom environment, teachers can apply specific instructional techniques much more successfully than in a larger classroom setting introduced by PD. For example, Lowellka et al. Again, (2019) describes group work as one of the strategies recommended in the PD program, and it seems reasonable to expect that such a strategy would involve a smaller number of co-participants.

Our analysis, therefore, demonstrates the existence of heterogeneous effects of the PD intervention, as well as identifies a wide range of potential determinants of variation [16]. Using GATES analysis, we show that the intervention hurts students in the bottom quintile while students in the top quintile are helped. This finding reinforces the argument by Lowellka et al. (2019): a segment of students benefits from the intervention, and a portion does not. Furthermore, our GATES analysis reveals that the impact on students in the middle quintiles is not statistically distinguishable from zero.

Through classification analysis, we are able to develop a more definitive profile of the characteristics that differentiate the students who benefit from the intervention from those who do not than what we had derived from the original HTE analysis. As the case was with Lowellka et al. (2019), teacher characteristics like having a college degree or majoring in mathematics could drive some variation. However, our study not only confirmed this finding but also revealed differential effects between those least and most impacted by the pandemic, not just in terms of baseline experience but also in terms of teacher rank, experience, age, and prior training, as well as student gender, age, baseline math anxiety, and class size, none of which were included in the original.

paper.

5. Conclusion

Our core message is that integrating predictive approaches with causal research questions enhances the value of traditional methods and should be explored more frequently in applied studies. We argue that researchers in each of the studies we revisited would have benefited from using causal machine learning (ML) techniques, gaining additional insights that standard causal inference tools do not provide.

For applied researchers, we offer the following recommendations regarding the implementation of causal ML approaches:

(a) Causal ML methods are particularly effective in settings with many potential confounders relative to the sample size. In our reviewed examples, we demonstrate the importance of accounting for all relevant confounders simultaneously, both linearly and non-linearly. We revisit the most comprehensive robustness checks from Jankow et al. (2010) and Nunn & Trefler (2010), where all potential confounders were considered both linearly and non-linearly—an approach that would not be feasible with traditional methods. Furthermore, our Monte Carlo simulations suggest that as the number of covariates included in the estimation increases relative to sample size, the advantages of using double machine learning (DML) over OLS become more pronounced.

(b) Causal ML approaches are better suited than traditional methods for flexibly capturing the effects of covariates. Since the true functional form is often unknown, using flexible estimation techniques allows for a more accurate representation of confounder effects. For instance, when revisiting the findings of Jankow et al. (2010), we present suggestive evidence of relevant nonlinear relationships that were overlooked in the original analysis but captured by DML estimation. Additionally, our Monte Carlo simulations demonstrate that in the presence of nonlinear confounders, DML outperforms OLS.

(c) We recommend using causal ML techniques in scenarios where researchers have limited theoretical guidance on which covariates should be included in the model. These methods provide a systematic approach to model selection, avoiding the ad hoc specification choices common in traditional approaches, as discussed in our reassessment of Jankow et al. (2010). This consideration is also crucial when conducting sensitivity analyses and robustness checks, as highlighted by our findings from revisiting Nunn & Trefler (2010).

(d) Lastly, when researchers are interested in heterogeneous treatment effects (HTE), causal ML methods can help ensure that relevant variation and its determinants are not overlooked or falsely identified due to multiple hypothesis testing issues. For example, our analysis of HTEs in studies by DellaVigna & Kaplan (2007) and Lowellka et al. (2019) reveals potential sources of variation that were not considered in the original analyses, which relied on traditional methods. Furthermore, causal ML approaches allow for the discovery of variation ex post rather than being restricted to subgroup analyses specified in a pre-analysis plan.

Compliance with ethical standards

Acknowledgments

The authors are grateful to the anonymous referees of journal for their extremely useful suggestion to improve the quality of the article.

Disclosure of conflict of interest

The author declared no potential conflicts of interest with respect to the research, authorship and publication of this article. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Funding

The author received no financial for the research, authorship and publication of this article.

References

[1] Athey, S. and G. W. Imbens (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 7353–60.

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

[3] Athey, S., J. Tibshirani and S. Wager (2019). Generalized random forests. Annals of Statistics 47, 1148–78. Athey, S. and S. Wager (2019). Estimating treatment effects with causal forests: An application. Observa- tional Studies 5, 37–51.

[4] Bertrand, M., B. Cre´pon, A. Marguerie and P. Premand (2017). Contemporaneous and post-program impacts of a public works program: Evidence from Coˆte d'Ivoire. Working paper, University of Chicago, IL. Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen and W. Newey (2017).

[5] Double/debiased/neyman machine learning of treatment effects. American Economic Review 107(5), 261–5.

[6] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. Econometrics Journal 21, C1–68.

[7] Chernozhukov, V., M. Demirer, E. Duflo and I. Fernandez-Val (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Working Paper 24678, National Bureau of Economic Research, Cambridge, MA.

[8] Colangelo, K. and Y.-Y. Lee (2020). Double debiased machine learning nonparametric inference with continuous treatments. arXiv: Econometrics 2004.03036.

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

[10] Davis, J. M. and S. B. Heller (2020). Rethinking the benefits of youth employment programs: The hetero- geneous effects of summer jobs. Review of Economics and Statistics 102, 664–77.

[11] DellaVigna, S. and E. Kaplan (2007a). The Fox News effect: Media bias and voting. Quarterly Journal of Economics 122, 1187–234.

[12] DellaVigna, S. and E. Kaplan (2007b). The Fox News effect: Media bias and voting [data]. Quarterly Journal of Economics. Data available at. https://eml.berkeley.edu/~sdellavi/index.html.

[13] Deryugina, T., G. Heutel, N. H. Miller, D. Molitor. and J. Reif (2019). The mortality and medical costs of air pollution: Evidence from changes in wind direction. American Economic Review 109(12), 4178–219.

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

[15] Djankov, S., T. Ganser, C. McLiesh, R. Ramalho and A. Shleifer (2010b). The effect of corporate taxes on investment and entrepreneurship [data]. American Economic Journal: Macroeconomics. Data deposited at ICPSR, https://www.openicpsr.org/openicpsr/project/114179/version/V1/view.

[16] Fair, R. C. (1978). The effect of economic events on votes for president. Review of Economics and Statistics 60, 159–73.

[17] Farrell, M. H., T. Liang and S. Misra (2021). Deep neural networks for estimation and inference. Econo- metrica 89, 181–213.

[18] Grossman, G. M. and E. Helpman (1991). Innovation and Growth in the Global Economy. Cambridge, MA: MIT Press.

[19] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20, 217–40.

[20] Imai, K. and M. Ratkovic (2013). Estimating treatment effect heterogeneity in randomized program evalu- ation. Annals of Applied Statistics 7, 443–70.

[21] Imbens, G. W. and D. B. Rubin (2015). Causal Inference in Statistics, Social, and Biomedical Sciences.

[22] New York: Cambridge University Press.

[23] Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation.

[24] Journal of Economic Literature 47, 5–86.

[25] Knaus, M. C., M. Lechner and A. Strittmatter (2022). Heterogeneous employment effects of job search programmes: A machine learning approach. Journal of Human Resources 57, 597–636.

[26] Kramer, G. H. (1971). Short-term fluctuations in us voting behavior, 1896–1964. American Political Science Review 65, 131–43.

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

[28] List, J. A., A. M. Shaikh and Y. Xu (2019). Multiple hypothesis testing in experimental economics. Experi- mental Economics 22, 773–93.

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

[30] Loyalka, P., A. Popova, G. Li and Z. Shi (2019b). Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program [data]. American Economic Journal: Applied Economics. Data deposited at ICPSR, https://www.openicpsr.org/openicpsr/project/116356/version/V1/view.

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

[32] Nunn, N. and D. Trefler (2010b). The structure of tariffs and long-term growth [data]. American Economic Journal: Macroeconomics. Data deposited at ICPSR, https://www.openicpsr.org/openicpsr/project/114183/version/V1/view.

[33] Oprescu, M., V. Syrgkanis and Z. S. Wu (2019). Orthogonal random forest for causal inference. Proceedings of the 36th International Conference on Machine Learning PMLR 97, 4932–41.

[34] Pissarides, C. A. (1980). British government popularity and economic performance. Economic Journal 90, 569–81.

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

[36] Strittmatter, A. (2019). What is the value added by using causal machine learning methods in a welfare experiment evaluation? Working paper, Global Labor Organization, Essen, Germany.

[37] Su, X., C.-L. Tsai, H. Wang, D. M. Nickerson and B. Li (2009). Subgroup analysis via recursive partitioning.

[38] Journal of Machine Learning Research 10, 141–58.

[39] Van der Laan, M. J. and S. Rose (2011). Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer Science and Business Media.

[40] Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113, 1228–42.

[41] Zeileis, A., T. Hothorn and K. Hornik (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics 17, 492–51.