American Economic Review: Papers & Proceedings 2017, 107(5): 546–550
https://doi.org/10.1257/aer.p20171000

LABOR MARKETS AND CRIME[‡]

Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs[†]

By JONATHAN M.V. DAVIS AND SARA B. HELLER*

Exploring treatment heterogeneity can provide valuable information about how to improve program targeting and what mechanisms drive results. But ad hoc searches for particularly responsive subgroups may mistake noise for a true treatment effect. Committing to a preregistered analysis plan can protect against claims of data mining or *p*-hacking but may also prevent researchers from discovering unanticipated results and developing new hypotheses. Modern corrections for multiple hypothesis testing are a useful alternative, though in practice they can be quite conservative when used across more than a few dimensions of heterogeneity.

Recent developments in the machine learning literature may help identify treatment heterogeneity in a principled way. Athey and Imbens (2016) and Wager and Athey (2015) extend regression tree and random forest algorithms to the problem of estimating average treatment effects for different subgroups, building on a large literature about estimating personalized treatment effects in medicine. If these methods adequately predict variation in treatment effects based on observable characteristics, they could provide more flexibility and better prediction of treatment heterogeneity by searching over high-dimensional functions of covariates rather than a small number of subgroups (typically defined by one or two interaction terms).

We apply Wager and Athey's (2015) causal forest algorithm to data from two randomized controlled trials (RCTs) of the same summer jobs program. Policymakers regularly make decisions about whom to serve in youth employment programs, so understanding how different subpopulations respond is substantively important in this context. Our setting is also a technically useful application because eligibility criteria were deliberately changed across the two RCTs to test for treatment heterogeneity; the studies have relatively large sample sizes for social experiments (1,634 and 5,216 observations, respectively); and we observe a large set of covariates.

Since this is an early application of the method, we focus our discussion on a step-by-step explanation of the process targeted at applied researchers. We explore how useful the predicted heterogeneity is in practice by testing whether youth with larger predicted treatment effects actually respond more in a hold-out sample. We conclude that although the method is likely to work best with datasets that are larger than our post-hold-out sample, it can identify treatment heterogeneity for some outcomes that typical interaction effects with adjustments for multiple testing would have missed.

[‡]*Discussants:* Bruno Crépon, CREST; Michael Mueller-Smith, University of Michigan; Stephen Raphael, University of California-Berkeley; Abigail Wozniak, University of Notre Dame.

*Davis: University of Chicago, 1129 E 59th Street, Chicago, IL 60637 (e-mail: jonmvdavis@gmail.com); Heller: University of Pennsylvania, 3718 Locust Walk, McNeil Building 483, Philadelphia, PA 19104 (e-mail: hellersa@upenn.edu). Research generously supported by B139634411 from the US Department of Labor, 2012-MIJ-FX-0002 from the Office of Juvenile Justice and Delinquency Prevention, and 2014-IJ-CX-0011 from the National Institute of Justice. We are deeply indebted to Susan Athey for providing an early beta version of the Causal Forest package. We thank Chicago Public Schools, the Department of Family and Support Services, the Illinois Department of Employment Security, and the Illinois State Police via the Illinois Criminal Justice Information Authority for providing data. The views expressed here are solely ours and do not represent the views of any agency or data provider.

[†]Go to https://doi.org/10.1257/aer.p20171000 to visit the article page for additional materials and author disclosure statement(s).

I. Background on Causal Forests

A standard regression tree algorithm predicts an individual's outcome, Y_i, using the mean Y of observations that share similar covariates, X. To define what counts as similar, the algorithm forms disjoint groups of observations, called "leaves," within which everyone shares values of certain Xs. A tree starts with the entire dataset in a single group. For every unique value of each covariate, $X_j = x$, the algorithm forms a candidate split of this group into two leaves by placing all observations with $X_j \leq x$ in a left leaf and all observations with $X_j > x$ in a right leaf. It implements the one split that minimizes an in-sample goodness-of-fit criterion such as mean squared error (MSE) $\left(\sum_{i=1}^{n} (\hat{y}_i - y_i)^2\right.$, where $\hat{y}_i$ is the mean Y within an observation's leaf$\left.\right)$. The algorithm then repeats the process for each of the two new leaves, and so on until it reaches a stopping rule. Using the last set of "terminal" leaves, the tree provides out-of-sample predictions by figuring out in which terminal leaf l an observation belongs based on its Xs and assigning $\hat{y}_i = \bar{y}_l$.

With a single tree, over-fitting is typically avoided by using a penalty parameter for the number of leaves selected via cross validation. But using a single tree is not always desirable; it is a high variance approach with no guarantee of optimality.[1] An alternative process called bootstrap aggregating selects hundreds or thousands of random subsamples of the data and grows a tree with no penalty in each subsample. Predictions of Y are the average of the $\hat{y}_i$s across all trees for individual i.[2]

In our case, however, we want to predict conditional average treatment effects (CATEs, or $E[Y_1 - Y_0 | X = x]$ in a potential outcomes framework) to assess how causal effects vary by subgroup. Standard fit measures like MSE are not feasible; unlike Y, $Y_1 - Y_0$ is not observed for any individual. Athey and Imbens (2016) introduce causal trees to solve this problem. They show that minimizing the expected MSE of predicted treatment effects, rather than the infeasible MSE itself, is equivalent to maximizing the variance of treatment effects across leaves minus a penalty for within-leaf variance. Within a tree grown using this modified criterion, CATEs are estimated as $\hat{\tau}_l = \bar{y}_{Tl} - \bar{y}_{Cl}$, or the treatment-control difference of mean outcomes within terminal leaf l. Here, $\hat{\tau}_l$ is the predicted treatment effect for out-of-sample observations with the Xs belonging to leaf l. To ensure correct inference, Athey and Imbens (2016) recommend an "honest" approach: divide the data in two, then use one subsample to determine the splits in the tree and the other subsample to estimate $\hat{\tau}_l$. Wager and Athey (2015) extend this idea to many trees and develop theory for inference in a causal forest (CF), which averages predictions from a large number of causal trees generated using subsamples of the full dataset.

II. Our Application

Our application uses two large-scale RCTs of Chicago's One Summer Plus (OSP) program conducted in 2012 and 2013. OSP provides disadvantaged youth ages 14–22 with 25 hours a week of employment, an adult mentor, and some other programming. Participants are paid Chicago's minimum wage ($8.25 at the time). The 2012 study ($N = 1{,}634$) is described in Heller (2014), which shows a 43 percent reduction in violent-crime arrests in the 16 months after random assignment. In 2013, we block randomized 5,216 applicants to OSP, 2,634 of whom were assigned to treatment.[3]

We match all applicants to administrative arrest records from the Illinois State Police (available for everyone), administrative schooling records from Chicago Public Schools (available for those who had ever enrolled in CPS), and unemployment insurance records from the Illinois Department of Employment Security (available for anyone who had a social security number in the CPS records, which was required for matching). There are no significant differences between the treatment and control groups' match rates. For this exercise, we focus on two outcomes of interest: the number

[1] Trees are typically built with "greedy" algorithms which choose the split that minimizes the MSE at a particular step, even if a different split would result in better predictive accuracy overall.

[2] This reduces bias by narrowing the neighborhood represented by each leaf and reduces variance by averaging across many predictions.

[3] Eligibility criteria changed to test program effects on a more criminally-involved population including only male youth. We individually randomized applicants within applicant pool, age, and geographic blocks.

of violent-crime arrests within two years of random assignment ($N = 6{,}850$) and an indicator for ever being employed during the six quarters after the program, defined only for those with non-missing employment data ($N = 4{,}894$).

III. Implementation Road Map

Using pooled data across both OSP RCTs, we implement a version of Wager and Athey's (2015) algorithm with a modification of their causalForest R package. The steps are as follows:

(i) Draw a subsample b without replacement containing $n_b = 0.2N$ observations from the N observations in the dataset.

(ii) Randomly split the n_b observations in half to form a training sample (tr) and an estimation sample (e) such that $n_{tr} = n_e = \frac{n_b}{2}$. Using *just* the training sample, start with a single leaf containing all n_{tr} observations.

(iii) For each value of each covariate, $X_j = x$, form candidate splits of the observations into two groups based on whether $X_j \leq x$. Consider only splits where there are at least ten treatment and ten control observations in both new leaves. Choose the single split that maximizes an objective function O capturing how much the treatment effect estimates vary across the two resulting subgroups, with a penalty for within-leaf variance (see the online Appendix for details and definition of O). If this split increases O relative to no split, implement it and repeat this step in both new leaves. If no split increases O, this is a terminal leaf.

(iv) Once no more splits can be made in step 3, the tree is defined for subsample b. Move to the estimation sample, and group the n_e observations into the same tree based on their Xs.

(v) Using *just* the estimation sample, calculate $\hat{\tau}_l = \bar{y}_{Tl} - \bar{y}_{Cl}$ within each terminal leaf. This step makes the tree honest, since treatment effect estimates are made using different observations than the ones that determined the splits.

(vi) Return to the full sample of N observations. Assign $\hat{\tau}_{l,b} = \hat{\tau}_l$ to each observation whose Xs would place it in leaf l, and save this prediction.

(vii) Repeat steps (i) to (vi) $B = 25{,}000$ times.

(viii) Define observation i's predicted CATE as $\hat{\tau}_i^{CF}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\tau}_{l,b}$, the average prediction for that individual across trees.

The procedure requires the researcher to select three parameters: the number of trees, the minimum number of treatment and control observations in each leaf, and the subsample size. In the absence of formal criteria to guide our choices, we used a large number of trees (more trees reduce the Monte Carlo error introduced by subsampling; we found moving from 10,000 to 25,000 improved the stability of estimates across samples). Increasing the minimum number of observations in each leaf trades off bias and variance; bigger leaves make results more consistent across different samples but predict less heterogeneity. Smaller subsamples reduce dependence across trees but increase the variance of each estimate (larger subsamples made little difference in our application).

For the CF to produce consistent estimates of program impacts, treatment assignment must be orthogonal to potential outcomes within each leaf (the "unconfoundedness" assumption). For this to be true in our case, we must condition on randomization block, since treatment probabilities vary across blocks. We adjust for differences in treatment probabilities using inverse probability weights throughout the procedure, including calculations of treatment effects and variances $\left(weight_i = \frac{T_i}{p_{block(i)}} + \frac{1 - T_i}{1 - p_{block(i)}}\right.$, where T_i is an indicator for being randomly assigned to the treatment group and $p_{block(i)}$ is the probability of being treated in observation i's block$\left.\right)$.

In defining covariates, we have to deal with missing data (e.g., prior-year wages are only available for those with valid SSNs who worked). To minimize missingness, we define mutually exclusive categories of covariates which are observed for all observations (e.g., a set of indicators for working in the year prior to

the quarter of randomization, not working in that year, or having missing employment data). In total, we use 19 covariates as inputs in the CF.[4]

IV. A Test of the Predictions

In the spirit of a standard subgroup analysis, but with subgroups determined by the high-dimensional combination of covariates captured by $\hat{\tau}_i^{CF}(x)$ rather than a few interactions, we ask: If we divide the sample into a group predicted to respond positively to OSP and one that is not, would we successfully identify youth with larger treatment effects? To test this, we first randomly split our 6,850 observations in half to create in- and out-of-sample groups, S_{in} and S_{out}. We run the entire CF procedure using only S_{in}, then use the trees grown in S_{in} to generate predictions for all observations in both samples.[5] This allows us to assess the performance of the predictions in a hold-out sample (albeit with reduced statistical power) and to check whether heterogeneity is more distinct in S_{in} than S_{out}, which could be a sign of overfitting.

Within each sample, we group youth by whether they are predicted to have a positive or negative treatment effect ($\hat{\tau}_i^{CF}(x) > 0$ is desirable for employment and adverse for arrests). We estimate separate treatment effects for these two subgroups by regressing each outcome on the indicator $\mathbf{1}[\hat{\tau}_i^{CF}(x) > 0]$, $T_i \times \mathbf{1}[\hat{\tau}_i^{CF}(x) > 0]$, $T_i \times (1 - \mathbf{1}[\hat{\tau}_i^{CF}(x) > 0])$, the baseline covariates used in the CF, and block fixed effects.[6] We then test the null hypothesis that the treatment effects are equal across the two subgroups. Rejecting the null would suggest that the CF predictions successfully sort the observations into two groups that respond differentially.[7] We do not adjust our inference for the fact that our regressors are defined by estimates themselves; calculating uniformly valid standard errors for CF predictions is still an open question. However, generating an indicator variable based on the predictions should reduce estimation error relative to using the predictions directly, since the error is less likely to matter for observations far from the zero threshold.

Table 1 shows the results. In panel A, which uses S_{in} only, the CF appears to identify distinct treatment heterogeneity for both outcomes. The subset of youth with predicted positive impacts shows a significant positive impact on average, the remaining youth have significant negative treatment effects, and the difference across these two groups is statistically significant. Panel B shows analogous estimates for S_{out}, where the difference in impacts between the predicted positive and negative responders is largely attenuated; we can reject the null that the subgroup difference is equal across S_{in} and S_{out} ($p < 0.01$ for both outcomes).

The difference between in- and out-of-sample results could either be because of an unlucky sample split or because there is some overfitting in S_{in}. To distinguish the two explanations, we make a small modification to the CF algorithm. In step (viii), instead of averaging across all trees to predict an individual's CATE, we only average across trees in which that observation was not part of either the tree-building or estimation samples.[8] Table 1, panel C shows results

[4] This is smaller than many "big data" settings but fairly standard for large social experiments. Other covariates are demographic characteristics (age in years and indicator variables for being male, Black, or Hispanic), neighborhood data (census tract unemployment rate, median income, and the proportions of tract with at least a high school diploma and who rent their home), education categories (indicator variables for having graduated from CPS prior to the program, being enrolled in CPS in the preprogram school year, not being enrolled in the preprogram year despite having a prior CPS record, and not being in the CPS data at all), criminal history (number of arrests at baseline for violent, property, drug, and other crimes), and the employment indicators described above. Gender is missing for 351 observations, which we impute using block means from the rest of the sample.

[5] We stratify the sample split by block, treatment status, and having a valid SSN, resulting in S_{in} having 3,428 observations. So each CF iteration uses 684 observations, and $n_{tr} = n_c = 342$.

[6] This estimates separate intent-to-treat effects for predicted positive and negative responders, capturing differences in both take-up rates and treatment responses across groups.

[7] There are many other ways to test whether there is useful information in the predictions. Variants of our test might make different cut-off decisions (e.g., serve the highest quartile of predicted responders or those whose CATEs are significantly different from 0 using the standard errors in Wager and Athey 2015), or interact $\hat{\tau}_i^{CF}(x)$ with treatment directly. Alternative tests could also address different questions entirely (e.g., how much of the variation in effects the predictions explain).

[8] That is, for an individual in S_{in}, we only average predictions from around 80 percent of iterations that did not include that observation in the 20 percent subsample; S_{out}

TABLE 1—TREATMENT EFFECTS BY PREDICTED RESPONSE

Subgroup	No. of violent crime arrests	Any formal employment
Panel A. In sample		
$\hat{\tau}_i^{CF}(x) > 0$	0.22	0.19
	(0.05)	(0.03)
$\hat{\tau}_i^{CF}(x) < 0$	−0.05	−0.14
	(0.02)	(0.03)
H_0: subgroups equal, $p =$	0.00	0.00
Panel B. Out of sample		
$\hat{\tau}_i^{CF}(x) > 0$	−0.01	0.08
	(0.05)	(0.03)
$\hat{\tau}_i^{CF}(x) < 0$	−0.02	−0.01
	(0.02)	(0.03)
H_0: subgroups equal, $p =$	0.77	0.02
Panel C. Adjusted in sample		
$\hat{\tau}_i^{CF}(x) > 0$	−0.06	0.05
	(0.04)	(0.03)
$\hat{\tau}_i^{CF}(x) < 0$	−0.02	−0.04
	(0.02)	(0.03)
H_0: subgroups equal, $p =$	0.41	0.02

Notes: Dependent variables are a count of violent-crime arrests over 2 post-random assignment years and an indicator for any formal employment over 6 post-summer quarters. Table shows intent-to-treat effects for youth whom the CF predicts will have a positive or negative treatment effect based on their covariates (standard errors clustered on individual). p-values from test that subgroup treatment effects are equal. Panels A and B show estimates for the sample used to estimate the CF ($N = 3{,}428$) and a hold-out sample which was not used to estimate the CF ($N = 3{,}422$), respectively. Panel C shows estimates for the same sample as panel A, but with an adjustment to avoid overfitting (see text for details).

using this adjusted approach, which are much more similar to the S_{out} results: we can no longer reject the null that the subgroup differences are the same in and out of sample ($p = 0.45$ and 0.99 for violence and employment, respectively). It seems, then, that including the observations used in tree-growing and estimation in the predictions generates some overfitting in our setting.[9]

With the adjustment, the subgroups split by $\hat{\tau}_i^{CF}(x) > 0$ no longer show significantly different treatment effects for violent-crime arrests. For employment, on the other hand, the predicted positive and negative responders have significantly different treatment effects in both the adjusted S_{out} and S_{in} samples. One caveat is that the success of the CF in identifying heterogeneity out-of-sample varies somewhat depending on how we split the sample (not shown). This sensitivity to the sample suggests that although our split sample is large relative to many other social experiments (over 3,400 observations in both S_{in} and S_{out}), the causal forest may be most useful in settings with more observations.

In our setting, the CF successfully identifies two subgroups with distinct employment effects. Standard interaction effects, on the other hand, fail to uncover heterogeneity that survives modern adjustments for multiple testing. The causal forest does not detect heterogeneity in violence impacts, which could happen for a few reasons. First, the treatment effects may not vary with observed covariates, either because unobservables drive treatment heterogeneity or because treatment effects are homogeneous. Second, the greedy algorithm may fail to identify the true functional form of the treatment effect, or our subgroup test may not isolate the true form of the heterogeneity. Finally, treatment heterogeneity could be obscured by sampling error; the CF may need bigger datasets.

REFERENCES

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences of the United States of America* 113 (27): 7353–60.

Heller, Sara B. 2014. "Summer Jobs Reduce Violence among Disadvantaged Youth." *Science* 346 (6214): 1219–23.

Wager, Stefan, and Susan Athey. 2015. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." https://arxiv.org/abs/1510.04342v2 (accessed February 23, 2016).

predictions are unchanged. This adjustment is a version of split-sample approaches; it completely separates estimation and prediction. If results still differ across S_{in} and S_{out}, the difference must be driven by something other than overfitting (e.g., the samples differ by chance). We note, though, that it is an ad hoc solution which may require adjustments to the CF's theoretical justification and inference.

[9] One way to reduce overfitting is to increase the minimum number of treatment and control observations in each leaf in S_{in}. A leaf size of 25 treatment and control observations reduces but not does eliminate the differential performance across samples, though it also does a worse job identifying heterogeneity out-of-sample than our adjustment.