AEA Papers and Proceedings 2021, 111: 440–444
https://doi.org/10.1257/pandp.20211090

Machine Learning about Treatment Effect Heterogeneity: The Case of Household Energy Use[†]

By Christopher R. Knittel and Samuel Stolper*

The rise of randomized controlled trials in economics has produced a wealth of evidence on the average causal effect of a great number of social and private sector programs. Yet such programs often have divergent impacts across the treated population. Understanding how different subgroups respond to a given treatment has the potential to unlock large increases in program effectiveness by allowing for improved targeting of the existing treatment (that is, identifying *whom* to treat) as well as improved design of the treatment itself (for example, tailoring treatment for specific subgroups).

Machine learning methods are an attractive option for identifying heterogeneous treatment effects (TEs) (Athey and Imbens 2016, Chernozhukov et al. 2018), because they offer tools for estimation that minimize the need for parametric assumptions and maximize out-of-sample predictive accuracy. In this paper, we estimate the heterogeneous TEs of a large-scale randomized experiment in household energy use. The treatment is the Home Energy Report (HER), a common behavioral nudge toward household energy conservation. We use the causal forest algorithm (Wager and Athey 2018) to predict TEs among 700,000 households and investigate the role of household characteristics in determining outcomes.

*Knittel: MIT Sloan School of Management (email: knittel@mit.edu); Stolper: University of Michigan School for Environment and Sustainability (email: sstolper@umich.edu). Leila Safavi and Paula Meloni provided outstanding research assistance. We thank Hunt Allcott, Tatyana Deryugina, Stefan Wager, and the many seminar participants that provided feedback. This research would not be possible without the work of Amy Findlay and colleagues at Eversource, who supplied the necessary data and background on the Home Energy Report Program.

[†]Go to https://doi.org/10.1257/pandp.20211090 to visit the article page for additional materials and author disclosure statement(s).

Our results contribute to an emerging empirical literature leveraging machine learning methods (Davis and Heller 2020, Kleinberg et al. 2018) as well as a large literature on the TEs of behavioral nudges (Ferraro and Price 2013; Andreoni, Rao, and Trachtman 2017). HERs in particular have been well studied: they reduce consumption on average (e.g., Allcott 2011), but there is some evidence of heterogeneity (Costa and Kahn 2013, Allcott and Kessler 2019), including boomerang effects (that is, *increases* in consumption; Byrne, Nauze, and Martin 2018). We build on this literature by predicting the full distribution of individual TEs of a widely used nudge as well as by identifying important correlates.

Our difference-in-difference estimate of the pooled average treatment effect (ATE) across all HER program waves is a reduction in monthly electricity usage of 9 kilowatt-hours (kWh), or 1 percent. The causal forest produces a full distribution of predicted individual TEs, ranging from -40 to $+10$ kWh, and with multiple statistical "modes" of response. In the first year of treatment, one mode is centered on -9 kWh, while another is centered on zero. In subsequent years, the modes diverge: the households that reduced consumption in year one ramp up their reductions, while boomerang effects become increasingly prevalent.

The most commonly used household characteristics in the forest are baseline (that is, pre-treatment) consumption and home value, which indicates that these variables in particular have significant predictive power. However, the bivariate relationships between individual treatment effect and each of these variables are not linear; the forest captures predictive effects that may not be apparent in the results of conventional regression models. In aggregate, the results of the causal forest indicate significant potential for efficiency improvements through selective targeting and adjustment of the HER "treatment."

I. Empirical Strategy

A. *Data*

We work with Eversource, an electric and natural gas utility in the northeastern United States, to study the impact of HERs. We collect data from households included in 15 waves of HER experiments that began between 2014 and 2017 (see Knittel and Stolper 2019 for further details). We observe treatment assignment, wave start date, monthly electricity consumption in kWh from 2013 to 2018, and 14 cross-sectional home and household attributes obtained from Eversource and Experian.[1]

We drop households with outlier values of home square footage and number of rooms as well as those enrolled in multiple HER waves or owning multiple properties. We further limit our sample to those households for which at least 12 months of preexperiment data and 12 months of postexperiment data are available. This leaves us with 902,581 households and a total of 35,959,282 household-monthly observations. For the causal forest, we fill in missing values of household characteristics using multiple imputation. Details on the multiple imputation procedure and summary statistics—including evidence of treatment-control balance in household characteristics—are provided in Knittel and Stolper (2019).

B. *Estimation of Average Treatment Effects*

We use our household-monthly panel data on electricity consumption to estimate wave-specific ATEs via the following regression:

$$(1)\quad kWh_{it} = \alpha_1 + \alpha_2 T_{it} + \mathbf{X}_i\eta + \boldsymbol{\theta}_i + \boldsymbol{\omega}_t + e_{it}.$$

Here, kWh_{it} is electricity consumption for household i in year-month t. The term T_{it} is the randomized binary treatment variable, $\mathbf{X}_i$ is a vector of household characteristics, and $\boldsymbol{\theta}_i$ and $\boldsymbol{\omega}_t$ are vectors of household and year-month fixed effects, respectively. We cluster standard errors by zip code. The term α_2 is our estimate of wave ATE in kWh per month. To obtain a single pooled estimate, we calculate the average of wave ATEs weighted by wave sample size.

C. *Causal Forests*

The causal forest algorithm (Athey, Tibshirani, and Wager 2019) is an adaptation of random forests (Breiman 2001) for the measurement of causal effects. Random forests are themselves an ensemble method applied to classification and regression trees (Breiman et al. 1984), which employ recursive partitioning to split a sample into subgroups that maximize heterogeneity across splits. A tree is a single run of recursive partitioning into subgroups (or "leaves"); a forest is a collection of trees, where each tree is grown from a randomly drawn (bootstrapped) subsample of the data.

We implement the algorithm using the generalized random forests (*grf*) R package. We grow 10,000 trees. For each one, we draw a random 50 percent sample to use and a random subset of characteristics to be considered in tree growth. We grow trees using "honest estimation" (Athey and Imbens 2016) so that the initial sample for each tree is split in half: one subset is used to grow the tree structure, and the other subset is used to estimate leaf ATEs.

Within-leaf ATE estimation in the *grf* package is implemented as a cross-sectional, difference-in-means comparison between treatment and control groups. To take advantage of our panel data structure, we define our dependent variable as the difference between average monthly electricity usage in year X of the relevant HER program wave (where $X \in 1,2,3$) and average usage in the year prior to wave start date. Additionally, and following Athey and Wager (2019), we "orthogonalize" our dependent and treatment variables by regressing each of these on observable characteristics and wave fixed effects (weighting observations by inverse probability of treatment) and recovering the residuals (see Knittel and Stolper 2019 for further details).

II. Results

Figure 1 displays ATE estimates in each individual program wave as well as for the full pooled sample. The pooled ATE is −9.4 kWh

[1] These are home age; home value; home square footage; number of rooms; age of household respondent; number of adult residents; income; educational attainment; an index for "green awareness"; average baseline consumption; and indicators for the presence of children, single-family occupancy, owner occupancy, and take-up of a subsidized home energy assessment.

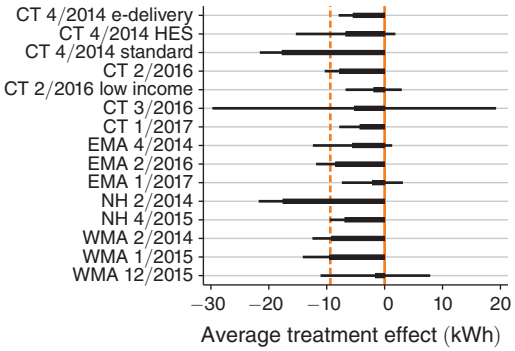

FIGURE 1. AVERAGE TREATMENT EFFECT BY WAVE

Notes: Each bar measures a wave-specific ATE. Error bars denote 95 percent confidence intervals. The vertical dashed line is the pooled ATE, measured as the average of wave ATEs weighted by wave sample size. CT = Connecticut; EMA = Eastern Massachusetts; NH = New Hampshire; WMA = Western Massachusetts.

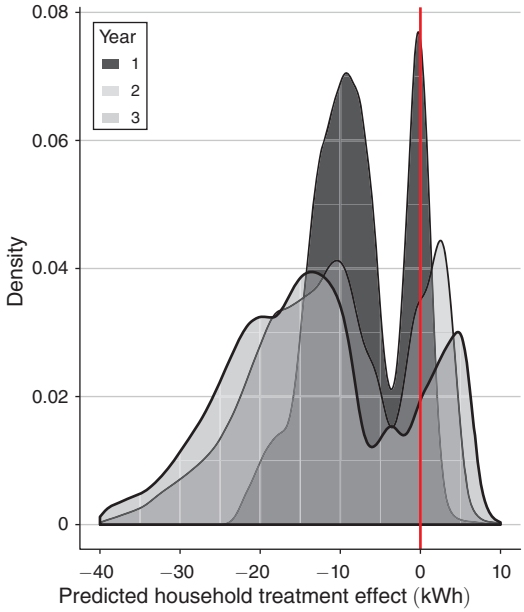

FIGURE 2. DISTRIBUTION OF PREDICTED TREATMENT EFFECTS

Notes: Each plotted distribution is a kernel density of household treatment effects in a specific year (one, two, or three) of HER programming. Treatment effect predictions come from our causal forest. The sample is fixed across years: only households with nonmissing consumption in all three post-years are included.

(per month), or −1 percent. Wave-specific ATEs range in magnitude from −1.6 to −17.7 kWh. The pooled ATE and 12 of the 15 individual program-wave ATEs are statistically significant at the 5 percent level or lower.

Figure 2 depicts the distribution of household treatment effect predictions produced by the causal forest. We plot separate distributions for each of the first three years of treatment. It is immediately clear from this graph that the distribution of TEs is multimodal. In year one of treatment, there is a large peak centered on −10 kWh as well as an even larger, albeit narrower, peak centered on zero. This zero peak implies that a significant number of households don't initially respond to, or perhaps even read, their HERs. In years two and three of treatment, both peaks progressively widen and shift away from zero. Households that initially respond by reducing consumption appear to learn to do more of that over time, but a sizable subset of the sample (18 percent) is predicted to *raise* its consumption. The full range of predicted TEs in year three extends from roughly −40 to +10 kWh.

What drives all this heterogeneity? Six characteristics—baseline consumption, home value, home square footage, the year in which a home was built, income, and respondent's age—are the most frequently used in the forest. Among these, the first two are easily the most common splitting variables. Baseline consumption is chosen as the initial splitting variable in 90 percent of trees in which it is eligible. Home value catches up to baseline consumption in frequency of use by the fourth split level. Beyond that point, these two attributes are used about twice as frequently as the other four (20 percent of the time versus 10 percent; see Knittel and Stolper 2019).

While frequency of use in tree growth provides some insight into the relative predictive power of characteristics, it does not clarify *how* these characteristics are related to TEs. To shed some light on these relationships, we zoom in on the two most frequently used characteristics: baseline consumption and home value. Figure 3 provides evidence on the relationship between the empirical distribution of predicted TEs and each of these two attributes. Each panel presents a scatterplot of individual values: the *y*-axis measures predicted TE, and the *x*-axis measures the attribute in question. We fit smooth, local polynomial functions to each scatterplot's data.

Both panels of Figure 3 hint at the potential value of nonparametric prediction methods such as the causal forest. Relatively simpler predictive

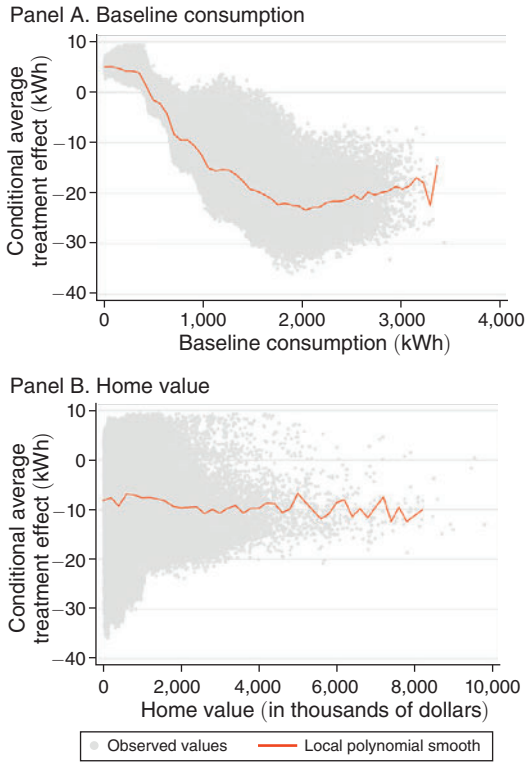

FIGURE 3. PREDICTED TREATMENT EFFECT VERSUS HOUSEHOLD TYPE

Notes: Each plotted point represents a household. In each panel, the *x*-axis measures the value of the indicated household characteristic, while the *y*-axis measures TE predicted by our causal forest. Lines depict the local smoothed polynomial relationship between ATE and the characteristic. The sample includes all households with nonmissing consumption in the year prior to program start and at least one of the first three years following program start. The dependent variable in the forest is average consumption across the three post-years minus average consumption in the first pre-year.

models may miss the nonlinearity of the relationship between treatment effect and baseline consumption, or they may miss the importance of the home value variable altogether. Panel A exemplifies the potential for improved program outcomes through selective targeting on observable characteristics. Setting the threshold for program inclusion around 800 kWh per month, for example, would be predicted to avoid nearly all boomerang effects. Meanwhile, if one wanted to better understand the characteristics of the very largest "reducers," panel B is helpful; such households are confined to the very bottom of the home value distribution. Nobody with home value above $100,000 is predicted to reduce monthly consumption by more than 23 kWh, while the households below that dollar threshold in some cases are predicted to reduce by 30–35 kWh.

III. Conclusion

Machine learning holds great promise as a tool for high-resolution evaluation and prediction. In this paper, we test that promise in the context of a large-scale experiment promoting household energy conservation. We leverage 15 experimental waves covering 700,000 households, in which the treatment is a periodic social comparison message designed to nudge households to reduce electricity consumption.

The causal forest that we estimate reveals several facts about TEs in this context. First, there is wide variation in responses to HERs. The overall average treatment effect is a 9 kWh monthly reduction in electricity consumption, but individual effects range from −40 to +10 kWh. Second, some households reduce more over time, while others tend toward *increases* in consumption. Third, baseline consumption and home value are the household characteristics most frequently used to grow the forest. Altogether, these facts illustrate the potential for improved targeting and tailoring of treatment through machine learning.

REFERENCES

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9–10): 1082–95.

Allcott, Hunt, and Judd B. Kessler. 2019. "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons." *American Economic Journal: Applied Economics* 11 (1): 236–76.

Andreoni, James, Justin M. Rao, and Hannah Trachtman. 2017. "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." *Journal of Political Economy* 125 (3): 625–53.

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

Athey, Susan, and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." Unpublished.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47 (2): 1148–78.

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

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* Boca Raton, FL: Routledge.

Byrne, David P., Andrea La Nauze, and Leslie A. Martin. 2018. "Tell Me Something I Don't Already Know: Informedness and the Impact of Information Programs." *Review of Economics and Statistics* 100 (3): 510–27.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." NBER Working Paper 24678.

Costa, Dora L., and Matthew E. Kahn. 2013. "Energy Conservation 'Nudges' and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment." *Journal of the European Economic Association* 11 (3): 680–702.

Davis, Jonathan M.V., and Sara B. Heller. 2020. "Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs." *Review of Economics and Statistics* 102 (4): 664–77.

Ferraro, Paul J., and Michael K. Price. 2013. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment." *Review of Economics and Statistics* 95 (1): 64–73.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics* 133 (1): 237–93.

Knittel, Christopher R., and Samuel Stolper. 2019. "Using Machine Learning to Target Treatment: The Case of Household Energy Use." NBER Working Paper 26531.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–42.