Predictive Counterfactuals for Treatment Effect Heterogeneity in Event Studies with Staggered Adoption

Mateus Souza*
Universidad Carlos III de Madrid

June 19, 2022

Abstract

I propose a machine learning (ML) based approach for estimating treatment effect heterogeneity in event studies with staggered adoption. The first step is to use ML algorithms to predict counterfactual outcomes in the absence of treatment, which can then be used to estimate a distribution of treatment effects. From this distribution it is possible to estimate different causal parameters according to the researcher's objective. With simulations in the context of high-frequency air pollution data, I show that the ML estimates are unbiased and more efficient than estimates from conventional approaches. My proposal serves as an alternative to standard two-way fixed effects regressions which, for example, have been shown to be near-term biased in the presence of dynamic treatment effects. I conclude with an application of the method to real data from a residential retrofit program, revealing substantial heterogeneity of energy savings depending on the types and levels of upgrades performed.

Keywords: Causal Inference, Machine Learning, Event Studies, Energy Efficiency, Air Pollution
JEL Classifications: C18, C55, Q49

*Address: Calle Madrid 126, Getafe, Madrid, Spain 28903; Email: mateus.nogueira@uc3m.es.
Acknowledgements: I thank Erica Myers, Peter Christensen, and Natalia Fabra for their guidance throughout this project. Camila Steffens also provided invaluable feedback. I am grateful for the support from Mick Prince and Chad Wolfe from the Illinois' Department of Commerce & Economic Opportunity, as well as Paul Francisco and Stacy Gloss from the Indoor Climate Research & Training center, without whom this research would not have been possible. Finally, I acknowledge the insightful comments from seminar participants at the University of Illinois, Universidad Carlos III de Madrid, ZEW Mannheim, Universidade Católica de Brasília, at the 2020 AERE Virtual Conference, and at the 10th Mannheim Conference on Energy and the Environment. Notwithstanding, all errors are my own. This research was possible in part thanks to the financial support from the Alfred P. Sloan Foundation, from CAPES (Coordination for the Improvement of Higher Education Personnel – Brazil), and from the European Research Council (ERC, under the European Union's Horizon 2020 research and innovation programme, Grant Agreement No. 772331).

1 Introduction

The recent surge in data availability is associated with both challenges and opportunities in modern research in environmental and energy economics. Thanks to remote sensing, smart meters, smartphone and internet applications, for example, researchers now face increasingly large volumes of complex information. On the one hand, as the quantity of observations and variables increases, so might processing time and computing memory required for analyses (Jin et al., 2015). On the other hand, with more information researchers can, for example, control for more confounders or test more competing hypotheses, which could potentially lead to more nuanced insights about a given topic (Ghanem and Smith, 2021; Coble et al., 2018). For that purpose, novel statistical methods and machine learning (ML) are becoming increasingly popular. New applications in economics and other social sciences, for example, show that these tools can be helpful to evaluate the effects of policy changes, field experiments, weather shocks, and others.[1] In this paper, I propose a machine learning-based approach to estimate heterogeneous effects, specifically for event studies with staggered adoption: settings in which observational units are exposed to a policy change/treatment at different points in time. I demonstrate the approach within two contexts: (1) for the evaluation of (simulated) policies to reduce air pollution (PM_{10}) concentrations; and (2) for estimating heterogeneous effects of a residential energy efficiency program.

A standard approach for treatment effect estimation within panel data settings would be to regress an outcome of interest on a treatment indicator variable (equal to one for unit-by-time observations exposed to treatment), plus unit fixed effects (e.g., indicators for each individual, household, or home in the sample) and time fixed effects (e.g., indicators for each hour, day, or month in the sample). The associated coefficient of the treatment variable is typically referred to as a Two-Way Fixed Effects estimator (henceforward denoted TWFE). The interpretation of TWFE is clear in settings with only two units (treated and control) and two time periods (before and after treatment):

[1] For reviews, see: Athey (2019); Storm, Baylis, and Heckelei (2019); Ghoddusi, Creamer, and Rafizadeh (2019); Weersink et al. (2018).

it represents the average treatment effect on the treated (ATT, formally defined in section 2). For that estimate to be valid, a necessary assumption is that the outcomes for treated and control units would have followed a same path (parallel trends) in absence of treatment (Angrist and Pischke, 2008). Nonetheless, even when that assumption holds, recent literature finds that the interpretation of TWFE may not be trivial for high-dimensional panel data settings with more than two units and more than two time periods, especially in the presence of time-varying (dynamic) treatment effects.[2]

Borusyak, Jaravel, and Spiess (2021) were among the first to show how the coefficient obtained from a "static" TWFE will be a weighted average of effects over time.[3] Further, some of the weights can be negative, especially for time periods long after treatment, such that the estimator can be biased towards near-term effects. Goodman-Bacon (2021) extends those results and provides a framework to decompose which time periods and groups of units (e.g., untreated, early, mid, or late adopters) contribute most to the estimate. The paper shows that estimation weights depend not only on the size of each of those groups, but also on the variance of treatment assignments. For example, a standard regression approach will assign more weight to portions of the sample with higher treatment variance (i.e., portions with substantial overlap between treatment and control units). For event studies with staggered adoption, the implication is that observations in the middle of the panel (mid-adopters) will receive greater weights relative to early- and late-adopters.[4] This can be problematic in the presence of treatment effect heterogeneity over time and across units, and to the extent that the researcher may not be particularly interested in an estimator that accentuates the effects of mid-adopters.

To overcome these issues, I propose an estimation approach that builds on recent advances in machine learning algorithms. With simulations, I show that the machine learning approach does not suffer from the biases related to standard TWFE. This paper

[2]See, for example: Baker, Larcker, and Wang (2022); Ghanem and Smith (2021); Borusyak, Jaravel, and Spiess (2021); Goodman-Bacon (2021); Kropko and Kubinec (2020); Imai and Kim (2020); Strezhnev (2018).

[3]This result was already present in an earlier version of the working paper, published in 2017 (Borusyak and Jaravel, 2017).

[4]Note that for event studies with staggered adoption both the beginning and the end of the panel have low treatment variance because the beginning of the panel contains few treated units, while the end of the panel contains few control units.

is related to a growing literature that proposes alternatives to the standard TWFE for settings with staggered adoption.[5] Sun and Abraham (2021) propose an extended TWFE specification that includes cohort-specific indicators interacted with indicators for time relative to the treatment event. This allows for heterogeneity over time and across cohorts. One key limitation is that their approach does not consider the inclusion of covariates. In contrast, the methods from Callaway and Sant'Anna (2021) and Wooldridge (2021) allow for *pre-determined* covariates. Callaway and Sant'Anna (2021) propose a doubly-robust estimator that relies on the estimation of counterfactual outcomes and of (treatment) propensity scores to re-weight the unit/time-varying effects. The approach is doubly-robust in a sense that it requires the correct functional form specification of only one of the two estimating equations. Wooldridge (2021) also proposes an extension of TWFE, in line with Sun and Abraham (2021), but adding pre-determined covariates under assumptions of linearity and via a Mundlak device (Mundlak, 1978).

In turn, my proposal outlined in this paper does not restrict treatment effect heterogeneity, and does not require re-weighting or estimation of auxiliary propensity scores. In particular, my proposal allows for the inclusion of *exogenous* covariates that can change over time. This is important for settings where covariates are known to affect the outcome of interest. When estimating the effects of interventions on energy consumption or air pollution, for instance, researchers typically control for weather conditions, such as temperature and precipitation, assumed to be exogenous to treatment assignment. Within these settings, the proposed method can be summarized as follows: first, using only pre-treatment data, build a model for flexible prediction of the full distribution of post-treatment counterfactual outcomes using machine learning tools;[6] second, subtract observed (true) post-treatment outcomes from the predicted counterfactuals to obtain a full distribution of treatment effects; third, summarize the treatment effects with (sub)sample averages, or by projecting them onto available covariates (to obtain

[5] Among others, see: Sun and Abraham (2021); Callaway and Sant'Anna (2021); Wooldridge (2021); Athey and Imbens (2022); Marcus and Sant'Anna (2021). Also, de Chaisemartin and D'Haultfoeuille (2020) present a general framework, not restricted to staggered adoption designs. For a survey, see Chaisemartin and D'Haultfoeuille (2021).

[6] A formal definition of "counterfactuals" is presented in Section 2.

conditional averages).

My proposal is closely related to independent work from Borusyak, Jaravel, and Spiess (2021), Liu, Wang, and Xu (2021), and Gardner (2021), who focus on direct "imputation" of counterfactuals estimated with fixed effects least squares regressions. Confirming their findings, I show that imputation approaches do not suffer from the problems of standard TWFE. I further show how machine learning algorithms can result in better prediction accuracy, as measured through root-mean-square errors, thus leading to more efficient estimation of treatment effects. One additional advantage of machine learning is that, when sufficient covariates are available, it can produce accurate predictions without the inclusion of unit fixed effects. This is useful for imputing counterfactuals even for units for which no pre-treatment data are available. Finally, some ML algorithms are agnostic about the functional forms and the importance of variables used for prediction. The predictive step will thus be less susceptible to researcher bias in model selection.

Concerning inference, I build on insights from Borusyak, Jaravel, and Spiess (2021), who propose a formula for clustered standard errors that are conservatively adjusted for errors in the first step (of counterfactual predictions). As an extension, I emphasize the role of *cross-validation*, which serves to reduce bias in the estimation of the errors from the predictive step.[7] I show that in-sample (as opposed to cross-validated/out-of-sample) residuals underestimate errors in the predictive step, which translates to higher risk of over-rejection of hypotheses of null treatment effects.

In Section 2, I conceptually define a few causal parameters that may be of interest, based on the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974). I introduce identifying assumptions under which my proposed estimands identify these causal parameters. In Section 3, I formalize my estimation approach, and propose tests for the identifying assumptions. In Section 4, I use simulations with semi-synthetic data to demonstrate the properties of my approach, compared to other imputation methods and standard TWFE. My simulations are performed with publicly-available data from air pollution monitors and weather stations across Spain (MITECO, 2020; AEMET, 2020).

[7] For a survey on cross-validation approaches, see Arlot and Celisse (2010).

These can be considered high-dimensional data in a sense that they are recorded daily and at 311 distinct locations. First, I show that an ML algorithm can provide reliable predictions of air pollution (PM_{10}) concentrations in this setting. Second, I show that my ML-based estimation approach is accurate for estimating simulated heterogeneous effects, and does not suffer from the established biases from standard TWFE. Third, I find that the ML approach is more efficient (has smaller standard errors), which can be especially important for estimating nuanced effects of programs/policies.

In Section 5, I further demonstrate my approach with an application to real data from the Illinois implementation of the Weatherization Assistance Program, which is a large residential retrofit program in the US. I have access to administrative and energy billing data from over 34 thousand homes served by the Program between 2006 and 2016. While previous evaluations focus on average effects (Fowlie, Greenstone, and Wolfram, 2018; Allcott and Greenstone, 2017; Zivin and Novan, 2016), I provide novel evidence on heterogeneity of energy savings from these types of programs. For example, I find that insulation measures are associated with substantial energy consumption reduction, while the effects of window replacements are close to zero. I also show that furnace repairs and re-tuning can increase energy consumption, providing evidence of a rebound effect in this context (Gillingham, Rapson, and Wagner, 2016). Conversely, full furnace replacements are among the highest energy-saving measures. Estimation of fine-scale heterogeneity also allows me to perform upgrade-specific cost-benefit analyses. With those, I find that, among a comprehensive suite of measures performed in these homes, only insulation upgrades are associated with positive net benefits, depending on lifespan and carbon price assumptions.[8] This does not imply that other measures should not be performed, as my analyses do not incorporate health, safety, and comfort benefits, for example. Nevertheless, my proposed method and the results documented here may help guide future efforts to identify and target high-savings measures or homes, in order to improve the cost-effectiveness of residential retrofit programs.

This paper contributes to a growing literature that demonstrates how machine learn-

[8]Longer lifespans and a larger social cost of carbon can lead other measures to be cost-effective as well.

ing tools can be leveraged for estimation of causal effects, with parallels to g-computation (Robins, 1986; Yu and van der Laan, 2002), double machine learning (Chernozhukov, Chetverikov, et al., 2017), targeted maximum likelihood estimation (van der Laan and Rubin, 2006; Balzer, Petersen, and van der Laan, 2016), synthetic control (Abadie, Diamond, and Hainmueller, 2010), and causal trees (Wager and Athey, 2018; Athey and Imbens, 2016). That body of work and recent applications (e.g., Allcott and Kessler, 2019; Prest, 2020; Miller, 2020; Burlig et al., 2020; Christensen et al., 2021; Abrell, Kosch, and Rausch, 2022) have highlighted some advantages of these novel methods, compared to standard impact evaluation: more efficient estimation, which allows for recovering more nuanced treatment effects; variable selection; potential bias reduction from explicit modelling of propensity of treatment; construction of robust comparison groups in settings where data for "pure controls" are not available; and potential for improved targeting of treatment assignments. The method I introduce is not only applicable to the analysis of interventions on air pollution and energy consumption. Rather, it can be applied generally to event studies with staggered adoption, as long as the identifying assumptions hold (see Section 2) and the data availability allows for robust prediction of counterfactuals (see Section 3).

2 Setup and Identification

Consider a panel data setting, with $i = 1, ..., I$ units (e.g., homes, household, individuals, or states) observed over time periods $t = 1, ..., T$. Let $Y_{i,t}$ denote an outcome of interest. Building on the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974), let there be two potential states for the outcome of interest: *treated*, $Y_{i,t}(1)$, in case unit i in time t has been exposed to some "treatment" (policy change, program, experiment etc.); or *untreated*, $Y_{i,t}(0)$, in absence of exposure. The effect of treatment for unit i at time t can then be defined as the difference between the outcomes at both potential states:

$$b_{i,t} = Y_{i,t}(1) - Y_{i,t}(0) \quad .$$

At a given point in time, a unit is either treated or untreated, such that, in practice, researchers can only observe outcomes at one of those two potential states. For the remainder of this paper, I use the term "counterfactuals" to refer to the outcomes at their alternative, *unobservable* state. In the proposed setting, units can be treated at different moments in time and they remain treated for all the subsequent periods. Let unit i be first treated at period $t = q_i$. Then i is untreated from $t = 1, ..., q_i - 1$ and treated from $t = q_i, ..., T$. Considering a binary treatment, and that all units' treatment regimes are observable, I define a variable $D_{i,t}$ equal to one for all unit-by-time observations that are exposed to treatment (i.e., $t \geq q_i$), and zero otherwise.[9] The sample may contain units that are "never treated," with $q_i = \infty$, such that $D_{i,t} = 0$ for those units during the full sample period (i.e., $T < \infty$).

Finally, let there exist a set of covariates $\mathbf{X}_{i,t}$, which can vary by units and over time, and which may affect the outcome of interest.[10] Note that the covariates are not assumed to be predetermined (i.e., they need not be fixed over time, or determined prior to the treatment), but must be not affected by the treatment status (see Assumption 3 below).

2.1 Causal Parameters of Interest

The elements presented above serve as building blocks for defining causal parameters of interest. The focus of this paper will be on recovering average treatment effects on the treated (ATT). Let the ATT be a function of $\mathbf{X}$ as follows:

$$ATT(\mathbf{X}) = \mathbb{E}[b_{i,t}|\mathbf{X}_{i,t}, D_{i,t} = 1] \quad .$$

Since $ATT(\mathbf{X})$ depends on covariates, it allows for heterogeneity of effects based on $\mathbf{X}_{i,t}$. However, researchers may be particularly interested in the average effects irrespec-

[9] This paper focuses on settings in which treatment can be considered binary. For a discussion on differences-in-differences with varying treatment "intensity," or continuous treatment, see Callaway, Goodman-Bacon, and Sant'Anna (2021).

[10] For ease of notation, throughout this paper I use **bold** print to indicate vectors such that, for example, $\mathbf{X}_{i,t} = [x^1_{i,t}, ..., x^K_{i,t}]$ represents K distinct covariates.

tive of covariates:

$$\begin{aligned} ATT &= \mathbb{E}[ATT(\mathbf{X})|D_{i,t}=1] \\ &= \mathbb{E}[b_{i,t}|D_{i,t}=1] \quad \text{, by the Law of Iterated Expectations (LIE).} \end{aligned} \tag{1.1}$$

The research might also be interested in the ATT for r periods of exposure relative to the treatment time, which is defined as:

$$ATT(r) = \mathbb{E}[b_{i,t}|D_{i,t}=1, t-(q_i-1)=r], \text{ for } r>0. \tag{1.2}$$

It may also be useful to get a sense of treatment effects for different subgroups or subsamples of the population, based on the covariates. Let $\mathbf{c}$ denote a set of conditions or rules on the covariates $\mathbf{X}_{i,t}$. Then a conditional average treatment effect on the treated (CATT) can be defined as:

$$CATT(\mathbf{c}) = \mathbb{E}[b_{i,t}|\mathbf{X}_{i,t}=\mathbf{c}, D_{i,t}=1] \ . \tag{2}$$

To recover the parameters described above, the fundamental problem of causal inference is that the untreated counterfactuals $Y_{i,t}(0)$ are not observable in post-treatment periods. I thus propose an estimation approach which requires, as a first step, the direct prediction of untreated counterfactuals, similar to the imputation method from Borusyak, Jaravel, and Spiess (2021). The key difference is that I allow for more flexible functional forms for the identification of the counterfactual, which relies on the following assumptions.

2.2 Identifying Assumptions

Assumption 1: *Random sampling.*

$$\{Y_{i,1}, Y_{i,2}, ..., Y_{i,T}, \mathbf{X}_{i,1}, \mathbf{X}_{i,2}, ..., \mathbf{X}_{i,T}, D_{i,1}, D_{i,2}, ..., D_{i,T}\}_{i=1}^{I} \tag{Asm. 1}$$

is independent and identically distributed (iid).

Assumption 1 states that the researcher has access to panel data. Note that it does not rule out time series dependence. Next, I assume that soon-to-be-treated units do not change their behavior in anticipation of treatment. The assumption of "no anticipatory effects" can then be formalized as:

Assumption 2: *No anticipatory effects.*

$$Y_{i,t} = Y_{i,t}(0) \text{ , for all } t < q_i \qquad \text{(Asm. 2)}$$

such that the observed pre-treatment outcomes are in fact outcomes in an untreated state. Further, I assume that covariates $\mathbf{X}_t$ are exogenous, such that:

Assumption 3: *Covariates are not affected by the treatment.*

$$\mathbf{X}_{i,t} = \mathbf{X}_{i,t}(0) = \mathbf{X}_{i,t}(1), \text{ for all } t \,. \qquad \text{(Asm. 3)}$$

Assumption 3 allows the inclusion of covariates that change over time, even in the post-treatment period.[11] Adding some structure, let there exist a function $g()$ that relates untreated potential outcomes with observable covariates $\mathbf{X}_{i,t}$, as follows:

$$\begin{aligned} Y_{i,t}(0) &= g(\mathbf{X}_{i,t}(0)) + \varepsilon_{i,t} \,, \\ \text{such that } \mathbb{E}[Y_{i,t}(0)|\mathbf{X}_{i,t}, D_{i,t} = 0] &= g(\mathbf{X}_{i,t}(0)) \,. \end{aligned}$$

Also let the conditional expectation of the treated potential outcome be given as follows:

$$\begin{aligned} \mathbb{E}[Y_{i,t}(1)|\mathbf{X}_{i,t}, D_{i,t} = 1] &= f(\mathbf{X}_{i,t}(1)) \,, \\ \text{where } Y_{i,t}(1) &= f(\mathbf{X}_{i,t}(1)) + \varepsilon_{i,t} \qquad . \end{aligned}$$

Allowing different functions for potential outcomes implies that there might be a change in the functional form of Y with respect to the covariates as a result of treatment. This warrants another key assumption, that, in expectation, the functional form of the

[11] In the absence of Assumption 3, for example, it would also be necessary to predict counterfactual realizations of covariates $\mathbf{X}_{i,t}(0)$ for $t \geq q_i$, which is out of the scope of this paper.

potential outcome would not have changed in the absence of treatment. I formalize this assumption below:

Assumption 4: *Stability of the counterfactual function.*

$$\mathbb{E}[Y_{i,t}(0)|\mathbf{X}_{i,t}, D_{i,t} = 1] = g(\mathbf{X}_{i,t}(1)) \; . \tag{Asm. 4}$$

Assumption 4 is analogous to the Conditional Parallel Trends assumption in the Difference-in-Differences literature. In the absence of treatment, the treated and control units would have followed a similar trajectory with respect to the covariates $\mathbf{X}$, which is determined by the function $g()$. This assumption implies that $g()$ can be used to understand the untreated counterfactuals also in the post-treatment periods.

Theorem 1: Under Assumptions 2 through 4, the ATT can be identified as follows.[12]

$$\begin{aligned} ATT &= \mathbb{E}[Y_{i,t} - g(\mathbf{X}_{i,t})|D_{i,t} = 1] \; , \\ ATT(r) &= \mathbb{E}[Y_{i,t} - g(\mathbf{X}_{i,t})|D_{i,t} = 1, t - (q_i - 1) = r], \; r > 0 \; . \end{aligned}$$

The identification of the $CATT(\mathbf{c})$ requires a more restrictive version of Assumption 4, namely:

Assumption 4': *Conditional stability of the counterfactual function.*

$$\mathbb{E}[Y_{i,t}(0)|\mathbf{X}_{i,t} = \mathbf{c}, D_{i,t} = 1] = \mathbb{E}[g(\mathbf{X}_{i,t}(1))|\mathbf{X}_{i,t} = \mathbf{c}, D_{i,t} = 1], \text{ for all } t \; . \tag{Asm. 4'}$$

Assumption 4' requires that, for each level of covariates $\mathbf{X}$, the counterfactual potential outcome for the treated is given, in expectation, by function $g()$.

Theorem 2: Under Assumptions 2 through 4', the $CATT(\mathbf{c})$ can be identified as follows.

$$CATT(\mathbf{c}) = \mathbb{E}[Y_{i,t} - g(\mathbf{X}_{i,t})|\mathbf{X}_{i,t} = \mathbf{c}, D_{i,t} = 1] \; .$$

[12] Assumption 1 is only required for consistency.

Proofs for Theorems 1 and 2 are provided in Appendix A. In the following section, I show how to estimate these causal parameters of interest using my proposed method.

3 Proposed Estimation Method

Within the setting described above, I propose an estimation approach based on direct prediction of untreated counterfactuals. The approach can be summarized in three steps: (i) building and selecting the predictive model; (ii) estimating the full distribution of treatment effects; and (iii) summarizing the estimated treatment effects. For the first step, I build on insights from the machine learning literature. I propose employing algorithms that allow for flexible relationships between the outcome and available covariates. Further, I show how cross-validation can be used for systematic model selection, and to help assess the validity of identifying assumptions.

3.1 Step 1: Building and selecting the predictive model

Given the assumptions and data structure described in Section 2, I propose the estimation of the treatment effects of interest based on the prediction of the counterfactuals $Y_{it}(0)$ for the treated units. The first step is to estimate the function $g()$, for which the researcher should only use the pre-treatment sample ($D_{i,t} = 0$).

In Economics, least squares regressions are often the method of choice. However, those often impose linear functional forms which may not be ideal for prediction accuracy, as I show later in Section 4. Rather, for this step I propose using more flexible approaches, such as tree-based methods, deep learners and neural networks, or support vector machines. Machine learning algorithms have been shown to outperform conventional approaches, especially with regards to *out-of-sample* prediction accuracy (Athey and Imbens, 2019; Varian, 2014). The properties of these algorithms are thus well aligned with the main objective of this first step: to predict counterfactuals that are in essence unobservable. To assess whether this objective has been met, it is key perform a careful and systematic analysis of out-of-sample prediction errors. This is standard within a

machine learning framework, and is typically done via cross-validation. In the context of this paper, cross-validation will also serve to assess the validity of the main identifying assumptions.

3.1.1 Cross-validation and model selection

In-sample and out-of-sample predictive performance of a given model may differ substantially. For example, the model may be excessively accurate (*overfitted*) for the sample in which it was trained, such that it cannot be generalized for other samples. This is problematic in the context of predicting counterfactuals that are, by definition, out-of-sample. I therefore propose cross-validation to systematically assess and compare accuracy of models being considered for this step. Cross-validation is also used for tuning hyperparameters,[13] and for defining the variables to be included, as well as their functional forms.

In this paper, I employ 5-fold cross validation as such:[14] (i) randomly split the pre-treatment sample into five equally sized subsamples; (ii) use four of these subsamples as the *training set* to estimate Y_{it} with a given model (and with a given set of hyparameter configurations), leaving one subsample aside as the *validation set*; (iii) using the model estimated in (ii), predict Y_{it} for the validation set; (iv) repeat steps (ii) and (iii) four times, such that all subsamples serve once as the validation set. It is then possible to obtain cross-validated predictions ($\hat{Y}_{it}^{cv}$) and residuals ($\hat{\varepsilon}_{i,t}^{cv} = Y_{i,t} - \hat{Y}_{i.t}^{cv}$) for the full sample.

The above process should be repeated for all the models and all the hyperparameter sets being considered. The researcher should then compare the validation set residuals

[13] *Hyperparameters* are set by the researcher prior to estimation, imposing some structure on the models being considered. Conventional *parameters*, are those estimated by the models. Examples of hyperparameters include: the maximum number regression trees for an ensemble; the minimum number of observations in the terminal nodes of the regression trees; the *learning rate* or weights associated with each new tree added to the ensemble.

[14] A different number of folds, or other cross-validation approaches may be considered, depending on the underlying data-generating process. For example, if the researcher is concerned about serial correlation in their setting, then they may apply some form of time series cross-validation (Hyndman and Athanasopoulos, 2018). If the researcher is further concerned about external validity, they may consider stratified subsampling such that, for example, the validation set never includes any observations from individuals in the training set. The best cross-validation approach will be context-specific, thus further guidelines on that are out of the scope of this paper. For a survey, see Arlot and Celisse (2010).

across all models, using metrics such as root-mean-square error (RMSE).[15] The algorithm and hyperparameter configurations that minimize cross-validated RMSE should be selected as the predictor for the function $g()$.

3.1.2 Assessing the identifying assumptions

Once the algorithm has been selected, the researcher can also assess whether Assumptions 2, 4, and 4' are likely to hold. Given that the counterfactual is not observed, assumptions 4 and 4' should be mostly based on economic knowledge (institutional, theoretical). In traditional difference-in-differences designs, researchers usually infer the validity of parallel trends based on the analysis of pre-treatment trajectories. As an alternative, within a machine learning framework I propose comparing predicted and observed outcomes (i.e., assessing the residuals from the predictive model).

Note that, by design, average residuals for most algorithms are very close to zero in the training set, but not necessarily in the validation set. The researcher should then first check if average cross-validated residuals are also close to zero. This will serve to assess if $\hat{g}()$ is accurate and stable for a new set of observations, thus providing insights on the validity of Assumptions 2 and 4. If cross-validated residuals are not close to zero, then there can be bias in Step 2 below, where prediction errors may be mistaken for treatment effects.

I propose a procedure that is similar to analyzing pre-trends in difference-in-differences settings. The researcher should run the following event study regression using pre-treatment periods only:

$$\hat{\varepsilon}^{cv}_{i,t} = \sum_{r \leq 0} \beta_r \mathbb{1}[r = t - (q_i - 1)] + u_{i,t} \text{ , for all } t < q_i \text{ ,} \tag{3}$$

where $\hat{\varepsilon}^{cv}_{i,t}$ are the cross-validated residuals; $\mathbb{1}[r = t - (q_i - 1)]$ are indicators equal to one for r periods relative to the treatment time, zero otherwise; and $u_{i,t}$ are idiosyncratic errors. The coefficients $\hat{\beta}_r$ will capture the average cross-validated residuals at r periods

[15] In this step, the researcher may choose to assess other error metrics as well, such as mean absolute error (MAE), R-squared, or others.

relative to treatment time. The researcher may graphically inspect $\hat{\beta}_r$ to check for any potential pretrends, and should perform an F-test to test if β_r are jointly zero, for all $r \leq 0$.

If the researcher is interested in heterogeneous treatment effects, then cross-validated prediction errors should be assessed for all the subsamples for which heterogeneity is expected. This is analogous to assessing the validity of Assumption 4'. If heterogeneity is expected to be a function of covariates, then I propose regressing:

$$\hat{\varepsilon}^{cv}_{i,t} = \boldsymbol{\beta}\mathbf{X}_{i,t} + v_{i,t} \text{ , for all } t < q_i \text{ ,} \tag{4}$$

where $\mathbf{X}_{i,t}$ are the covariates along which potential treatment effect heterogeneity is expected. The variables to be included, and their functional forms, should be determined by the researcher, according to their prior knowledge of the field, or depending on which dimensions of heterogeneity seem particularly interesting. The researcher may visually inspect $\hat{\boldsymbol{\beta}}$ for potential patterns in the errors, and should test if $\boldsymbol{\beta}$ are equal to zero. Essentially, the researcher should be able to show supporting evidence that cross-validated prediction errors are uncorrelated with the covariates that may drive treatment effect heterogeneity.[16]

In Section 4 and in Appendix D, I present details about the cross-validation results in a setting where the main target is to predict counterfactual air pollution concentrations. I compare results from both training and validation set prediction errors for algorithms considered in this paper. Further, in the context of simulations I am able to compare estimated errors to "true" counterfactual prediction errors (this is impossible in real data settings, given that the counterfactuals are never observed). I find that validation set residuals, as opposed to in-sample residuals, are a better proxy for the true prediction errors. I also find, for the real data application (Section 5), that my chosen prediction algorithm produces residuals that are not significantly correlated with any of the included covariates.

[16] The same independent variables from equation (4) will be used for summarizing the treatment effects in Step 3 below.

3.2 Step 2: Estimating the full distribution of treatment effects

For this second step, the researcher should evaluate the estimated function $g()$ at $\mathbf{X}_{i,t}$ for the treated units in the the post-treatment sample as follows:

$$\hat{Y}_{i,t}(0) = \hat{g}(\mathbf{X}_{i,t}) \text{ , for all } D_{i,t} = 1 \text{ .}$$

If step one was performed correctly, then it is possible to obtain the full distribution of $\hat{Y}_{it}(0)$ (i.e., for all treated units and for all post-treatment periods). Recall that those predictions are based on a model built with pre-treatment observations only, such that they can be viewed as counterfactual predictions for the treated units.

Unit-by-time treatment effects can then be estimated by:

$$\hat{b}_{i,t} = Y_{i,t} - \hat{Y}_{i,t}(0) \text{ , for all } D_{i,t} = 1.$$

which gives us the full distribution of treatment effects. $\hat{b}_{i,t}$ are the building blocks necessary to obtain the sample analogue of the causal parameters of interest described in Section 2, and whose estimation is described in the next step.

3.3 Step 3: Estimating the causal parameters of interest

The most prominent parameter of interest, especially in economics, is the average treatment effect on the treated (ATT), which can be estimated in this setting as:

$$\widehat{ATT} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}\hat{b}_{i,t}\mathbb{1}\{D_{i,t}=1\}}{\sum_{i=1}^{I}(T-(q_i-1))\mathbb{1}\{q_i \leq T\}} \text{ ,}$$

which is an average of all $\hat{b}_{it}$ obtained in Step 2. Alternatively, ATTs can be summarized as percentages, dividing by the sample average predicted counterfactual:

$$\%\widehat{ATT} = \widehat{ATT} \Bigg/ \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}\hat{Y}_{i,t}(0)\mathbb{1}\{D_{i,t}=1\}}{\sum_{i=1}^{I}(T-(q_i-1))\mathbb{1}\{q_i \leq T\}} \text{ ,}$$

where, again, all elements necessary for the above parameter have already been obtained

from steps 1 and 2. Consistency properties are discussed in Appendix A. Note that the above expressions for both $\widehat{ATT}$ and $\%\widehat{ATT}$ weight post-treatment observations equally. However, the researcher may choose to calculate these averages with different weights when, for example, data come from a survey where each individual represents a known portion of the population. For a discussion on when alternative weights are appropriate, see Solon, Haider, and Wooldridge (2015).

For describing heterogeneity in effects, it can be useful to calculate the above average parameters for different subsamples. For example, if the researcher is interested in heterogeneous effects depending on time of exposure to the treatment, then the ATT(r) can be estimated as follows:

$$\widehat{ATT}(r) = \frac{\sum_{i=1}^{I} \hat{b}_{i,t} \mathbb{1}\{t - (q_i - 1) = r\}}{\sum_{i=1}^{I} \mathbb{1}\{t - (q_i - 1) = r\}} \text{ , } r > 0 \text{ .}$$

In addition, a researcher might hypothesize that a certain subpopulation experiences treatment effects that are different than the ATT. If covariates allow that subpopulation to be identified, then the researcher can estimate:

$$\widehat{CATT}(\mathbf{c}) = \frac{\sum_{i=1}^{I} \sum_{t=1}^{T} \hat{b}_{i,t} \mathbb{1}\{D_{i,t} = 1\} \mathbb{1}\{\mathbf{X}_{i,t} = \mathbf{c}\}}{\sum_{i=1}^{I} (T - (q_i - 1)) \mathbb{1}\{q_i \leq T\} \mathbb{1}\{\mathbf{X}_{i,t} = \mathbf{c}\}} \text{ ,}$$

where $\mathbf{c}$ denotes a set of conditions on $\mathbf{X}_{i,t}$ that identify a subpopulation of interest. The $\%\widehat{CATT}$ can also easily be obtained by imposing conditions on $\%\widehat{ATT}$. Note that these are simply conditional averages of the treatment effects. For continuous covariates, c is a range of values that the covariate X can assume, which needs to be defined parsimoniously such that a large number of observations is found in $X = c$ and Assumption 4' holds.

If the researcher believes that treatment effects exhibit a more complex structure, then a linear regression can be considered:

$$\hat{b}_{i,t} = \boldsymbol{\beta} \mathbf{X}_{i,t} + u_{i,t} \text{ , for all } D_{i,t} = 1 \text{ ,} \tag{5}$$

where $u_{i,t}$ is an idiosyncratic error term; and $\boldsymbol{\beta}$ captures the relationship between *treat-*

ment effects and covariates of interest. The flexibility of that relationship will depend on the functional forms and interactions of $\mathbf{X}_{i,t}$. Note that the linear regression equation 5 is simply used to obtain conditional averages, and alternative methods may be considered. Further, identification still relies on the assumptions described in Section 2.

3.4 Inference

To obtain standard errors for each of the estimators described above, I follow a conservative approach from Borusyak, Jaravel, and Spiess (2021). Specifically, I propose a slight modification of their Theorem 3:

$$\hat{\sigma}^2_{cv} = \sum_i \Big(\sum_{t;D_{i,t}=0} \gamma_{i,t} \hat{\varepsilon}^{cv}_{i,t} + \sum_{t;D_{i,t}=1} \gamma_{i,t} \tilde{\varepsilon}_{i,t} \Big)^2 , \tag{6}$$

where I refer to $\hat{\sigma}^2_{cv}$ as the *cross-validated* variance; $\hat{\varepsilon}^{cv}_{i,t}$ are cross-validated residuals in the pre-treatment sample; $\tilde{\varepsilon}_{i,t} = \hat{b}_{i,t} - \hat{\bar{b}}_{i,t}$ are deviations of the estimated treatment effects ($\hat{b}_{i,t}$) from average effects ($\hat{\bar{b}}_{i,t}$); and $\gamma_{i,t}$ are sampling weights.[17] From Equation (6), it is possible to obtain standard errors clustered at the unit level, thus to perform inference according to the hypotheses to be tested.

The first term from the right-hand side of (6) is intended to capture errors from Step 1 (the predictive step). Different from Borusyak, Jaravel, and Spiess (2021), I propose using cross-validated residuals for that term, rather than in-sample residuals. This is because, as discussed in Section 3.1, cross-validated residuals are likely more accurate proxies for the true counterfactual prediction errors. Even under this conservative adjustment, I show that my proposal estimates treatment effects more efficiently compared to alternative approaches.

The choice of $\hat{\bar{b}}_{i,t}$ is up to the researcher, who should keep in mind a tradeoff between consistency and the size of $\hat{\sigma}^2_{cv}$. On the one hand, a conservative choice would be to take

[17] For the applications in this paper, all observations are weighted equally. This may not be ideal for all settings. The proposal from Borusyak, Jaravel, and Spiess (2021) involves calculating weights for efficient estimation of ATT. See Solon, Haider, and Wooldridge (2015) for a discussion on when and how to use alternative sampling weights.

a single average $\hat{\bar{b}}_{i,t}$ for the full post-treatment sample.[18] Note that for $\widehat{CATT}(\mathbf{c})$ the variance in (6) needs to be estimated for each group of observations defined by the set of conditions c. In that case, the averages for $\hat{\bar{b}}_{i,t}$ may be taken separately for these groups for which heterogeneity is expected. The key is that the subsamples should be large enough to retain consistency.

For cases where a complex heterogeneity structure is expected, as in equation (5), I recommend bootstrapping as a yet more conservative approach for estimating standard errors. Chernozhukov, Fernández-Val, and Luo (2018) show that bootstrapping can improve estimates of confidence bands in settings with substantial heterogeneity. The proposed bootstrap algorithm is presented in Appendix B, where I also show that the stability of estimated standard errors and the optimal number of bootstrap iterations are context specific.

4 Simulations

I demonstrate the method proposed in Section 3 with simulations using semi-synthetic data. That is, I use real-world data as the basis of my simulations but change some of the observed outcomes depending on each simulation's objectives. The outcome of interest is ambient particulate matter (PM_{10}) concentrations measured at 311 air quality monitors across Spain (MITECO, 2020).[19] Particulate matter concentrations are recorded daily. I use observations from the 1st of January 2014 to the 31st of December 2019. I match those with daily weather data from 271 stations across the country.[20] These weather stations provide the following key control variables: wind direction, wind speed, atmospheric pressure, precipitation, min, max, and median temperatures (AEMET, 2020).

I have additionally collected the following variables that might be related to air pollution concentrations: national-level daily electricity generation by fuel type (ESIOS, 2020);

[18] Note that the variance in equation (6) is inflated by treatment effect deviations from the calculated average effect $\hat{\bar{b}}_{i,t}$.

[19] I focus on monitors located in urban or suburban areas, dropping those in rural areas.

[20] Each air quality station is matched with the nearest weather station, based on simple linear distances between stations' coordinates.

province-level annual GDP, population, and employment (INE, 2020); province-level monthly entry and exit of firms (INE, 2020); and national-level annual hectares of forest area burned by wildfires (MITECO, 2021). Finally, the ML algorithms also include seasonality controls (year, month, and day of year FE, holidays, and a monthly trend), as well as characteristics from the air quality monitors (station type; altitude; urban or suburban location; industrial, commercial, or residential location). Descriptive statistics for the outcome variable and the full set of included controls are presented in Appendix C.

Using these data, I simulate scenarios consistent with the focus of this paper (i.e., staggered adoption). First, I assign random "artificial" treatment dates to each air quality station, thus allowing the identification of pre and post-treatment observations. Then, depending on the illustrative intent of each simulation, I impose a simulated treatment effect by changing the outcome (air quality concentrations) for post-treatment observations, leaving pre-treatment data unchanged. The outcome change (simulated effect) constitutes of lowering the PM_{10} concentrations after treatment, by subtracting a given percentage of the original value.[21] With this setup, I observe both simulated treatment effects and "ground truth" counterfactuals that can be used to assess performance of different estimation techniques.

For Step 1 of my proposed approach, I employ a machine learning algorithm called XGBoost, which is a computationally efficient implementation of gradient boosted trees (Chen and Guestrin, 2016). I perform ML prediction using the pre-treatment sample, defined based on the artificially allocated treatment dates. In Appendix D, I present performance metrics for XGBoost, which exhibits high cross-validated prediction accuracy in this setting.[22] As discussed in Section 3, cross-validation is essential for this step, given that I aim to predict counterfactuals which are unobservable by definition. To Illustrate this point, Figure 1 Panel A compares the distributions of *in-sample* versus true counterfactual residuals, while Panel B compares *cross-validated* versus true counter-

[21] This can be thought of as the effects of implementing low emission zones in cities, or mandating the installation of scrubbers in industrial facilities, for example.

[22] I use root-mean squared error (RMSE) as a measure of prediction accuracy. However, other metrics, such as R-squared or mean absolute error (MAE), may also be considered.

factual residuals. Note that in-sample and cross-validated residuals were obtained with pre-treatment data only, while true counterfactuals were obtained with post-treatment data. A comparison of Panels A and B reveals that in-sample residuals, as expected, are relatively better centered around zero, compared to cross-validated residuals. However, Panel B shows that cross-validated residuals exhibit better overlap with the true counterfactual residuals. The implication is that in-sample residuals would underestimate the errors from the predictive step. This highlights why cross-validated residuals are more appropriate for testing the identifying assumptions (as proposed in section 3.1.2), as well as for adjusting the variance for inference in equation (6).

[FIGURE 1 HERE]

In real data settings, however, a comparison between cross-validated and true counterfactual residuals is not feasible, since true counterfactuals cannot be observed. Alternatively, a researcher may choose to assess residuals in a separate "test" sample, which was not used for model selection or tuning. The assumption is that this "test" sample will be free from any biases introduced during the tuning process, and will accurately represent predictive performance for a completely new set of observations. The procedure for appropriately defining a test sample will be context specific, depending on the underlying data-generating process. For further discussion how to define training, validation, and testing samples, see Arlot and Celisse (2010).

I also assess no anticipatory effects and the stability of my estimated counterfactual function by regressing cross-validated residuals on indicators of time relative to treatment, as shown in equation (3). This is analogous to a "pre-trends" test in traditional difference-in-differences settings. Coefficient estimates and 95% confidence intervals from this regression are shown in Figure 2. Note that only one of the coefficients may be considered statistically significant (coefficient for 10 months prior to treatment). However, with an F-statistic of 1.11 (top right corner of the Figure) for a test of joint significance of the coefficients, I cannot reject that all the coefficients are jointly equal to zero, thus providing supporting evidence for Assumptions 2 and 4.

[FIGURE 2 HERE]

I next proceed by applying Steps 2 and 3 outlined in Section 3 to recover simulated treatment effects. I compare the performance of my proposed ML approach to standard TWFE, and to an imputation method that uses ordinary least squares for the predictive step (Step 1). Note that I consider several simulation scenarios, described in detail below, to highlight different features of my proposed approach. The predictive step of my approach, however, will remain the same for all simulations presented in this paper.

4.1 Treatment Effect Variation Across Time

For these simulations, I impose effects that are either increasing or decreasing over time. The rationale is to verify if my method is robust to dynamic treatment effects, in contrast to standard two-way-fixed effects regressions, which have been shown to suffer from near-term bias (e.g., Goodman-Bacon, 2021). For these analyses, first I restrict the sample such that each air quality station will have no more than 2 years (24 months) of data before and after treatment. I also restrict the sample to observations for which outcomes Y_{it} and covariates $\mathbf{X}_{it}$ are always jointly observable.[23] One implicit assumption is that observations are missing at random, such that missingness is orthogonal to treatment and other relevant factors. Then I impose simulated treatment effects by reducing post-treatment PM_{10} concentrations by a given percentage.

I start by simulating effects that decrease over time. I impose that PM_{10} reductions will be 20% for the first semester after treatment, 15% for the second semester, 10% for the third semester, and 5% for the fourth semester. In the context of pollution abatement policies, this can be viewed as a simulation with abatement technology depreciation over time. I then proceed by comparing ML estimates with the "ground truth" simulated

[23] That is necessary for the algebra of regression analyses: the number of rows of the outcomes' vector must be equal to the number of rows of the covariates' matrix. Researchers typically impose that by dropping rows with missing observations for a few key variables, which is the strategy that I also employ in the application of this paper.

effects, as well as with standard two-way fixed effects (TWFE) specifications such as:

$$Y_{i,t} = \beta \times D_{i,t} + \alpha_i + \alpha_t + u_{i,t} \ , \tag{7}$$

where $D_{i,t}$ is equal to one if air quality station i has been treated in day of sample t, zero otherwise; β is the TWFE parameter of interest, which captures a weighted average of the dynamic treatment effects; α_i and α_t are station and day of sample fixed effects, respectively; and $u_{i,t}$ is the error term. I also try variations of the above TWFE specification by interacting station and calendar month fixed effects, by including day of sample by province fixed effects, and by adding time-varying controls (i.e., weather variables).

My proposed ML estimates are also compared to those from a simpler imputation method, following Borusyak, Jaravel, and Spiess (2021), that uses ordinary least squares for predictions of counterfactuals. For that, I use equation (7) above, restricted to the pre-treatment sample, to estimate a model for the counterfactuals. For this simpler imputation approach, standard errors are adjusted based on in-sample, rather than cross-validated residuals.

Results for all specifications are presented in Table 1 Panel A. Column (1) presents the "true" effect which serves as the benchmark. Column (2) presents results from my ML approach. Columns (3) and (4) are for standard TWFE approaches. Columns (5) and (6) are for an imputation approach that uses OLS for prediction of untreated counterfactuals. I present the estimated effects as well as the standard errors (in parentheses) according to each approach. Standard errors from standard TWFE are clustered by air quality station. Standard errors for the ML approach are also clustered by station and are further adjusted using cross-validated residuals from the predictive step, according to equation 6. Finally, for comparison, conservative bootstrapped standard errors for the ML approach are presented in square brackets.

Results suggest that, compared to the ground truth, the machine learning approach provides an accurate estimation of the ATT. On the other hand, the coefficient obtained from standard TWFE overestimates the true savings by about 54%. Adding finer-scale

fixed effects and time-varying controls does not seem to significantly improve the accuracy of standard TWFE. Consistent with results from Borusyak, Jaravel, and Spiess (2021), the OLS imputation approaches outperform standard TWFE both in terms of bias and efficiency. However, standard errors from OLS imputation are about 60% larger than those obtained with the ML approach. One additional advantage of the ML approach is that it allows the researcher to retain more observations in the sample. That is because the ML predictive step does not include fine-scale fixed effects which, for the other approaches, need to be available in both pre- and post-treatment samples.

For Table 1 Panel B, I repeat the exercise above, but with treatment effects that are increasing over time. Now I impose that PM_{10} reductions will be 5% for the first semester after treatment, 10% for the second semester, 15% for the third semester, and 20% for the fourth semester. For this case, I find that standard TWFE underestimate the true effect. Taken together, Panels A and B show that standard TWFE exhibit a near-term bias in this setting, as suggested in prior literature. This bias, however, is not present in the ML and OLS imputation methods.

[TABLE 1 HERE]

Returning to the simulation with effects that decrease over time, I aim to test the performance of estimators for recovering ATT(r), which are the effects at given semesters after treatment. For TWFE, this can be estimated with a variant of equation (7) that includes a interactions of $D_{i,t}$ with indicators for semesters post-treatment. For the ML and OLS imputation approaches, ATT(r) can be estimated by taking averages of the unit-by-time effects ($\hat{b}_{i,t}$) over each of the semesters of interest. Inference for ML and OLS imputation procedures is also based on the adjustment shown in equation (6), but with each element calculated within the subsamples determined by each semester.

Results for ATT(r) are shown in Table 2. It can be noted that all the approaches provide unbiased estimates of the effects across semesters. This suggests that TWFE should not be dismissed for settings in which heterogeneity is only expected across time (i.e., when heterogeneity over other covariates is expected to be limited). For those cases,

a fully dynamic specification of TWFE can be unbiased, as shown in prior literature (e.g., Sun and Abraham, 2021). I highlight, however, that the ML approach exhibits substantial gains in efficiency, compared to the other methods. Note that, for all the semesters, the standard errors from the ML approach are smaller than those from the other methods. Next I assess the performance of these methods under treatment effect heterogeneity both across time and by other observable characteristics ($X_{i,t}$).

[TABLE 2 HERE]

4.2 Treatment Effect Heterogeneity Across Time and by Observable Characteristics

For this simulation, I consider a more complex treatment structure. Following what was done for the above simulations, I impose effects that decrease over time. Additionally, I use a covariate (altitude of air quality stations) to non-randomly split all the observations into four groups which will get different "bonus" treatment effects. As shown in column (1) of Table 3, stations with altitude below 35 meters get the strongest bonus effect, with resulting average PM_{10} reduction of about 7 $\mu g/m^3$. Stations with altitude between 35 and 150 meters have an average PM_{10} reduction of about 4.16 $\mu g/m^3$. Stations with altitude between 150 and 500 meters have an average PM_{10} reduction of about 1.2 $\mu g/m^3$. Stations with altitude higher than 500 meters experience no PM_{10} reductions.[24]

Now suppose that the researcher is interested in estimating treatment effects for each of the groups defined above, but is not particularly interested in (or chooses to ignore) how effects change over time. They should thus aim to estimate CATT, as defined by equation (2). For this, it is first necessary to test if the identifying assumptions hold for each group, by implementing the regression specification (4). As such, in Appendix Figure D.1, I show that the ML cross-validated prediction errors are not correlated with bins of altitude of the air quality stations.

[24]Note that I create four groups of stations based on their actual measures of altitude. For this illustrative simulation, I could have picked any other variable that satisfies the assumptions laid out in section 2.2.

I then proceed to estimate CATT with different approaches. Results are presented in Table 3. As for the case of (unconditional) ATT, it is clear that standard TWFE estimates are biased, with the coefficients being more representative of the near-term effects (that occur right after treatment). In contrast, ML and OLS imputation more accurately estimate the extent to which the stations were affected over the full two years after treatment. OLS imputation, however, exhibits bias in estimating the effects for the last altitude bin: the point estimate is close to a PM_{10} *increase* of 1 $\mu g/m^3$, while the benchmark effect is actually zero. The ML approach does not exhibit such bias, and further retains the advantages described in the previous section: more efficient estimation and no loss of observations.

[TABLE 3 HERE]

I next turn to an application of the ML approach for estimating CATT using real data from the Weatherization Assistance Program.

5 Real Data Application: Heterogeneous Effects of the Illinois Weatherization Assistance Program

Primary data from this application comes from the Illinois implementation of the Weatherization Assistance Program (WAP). WAP is a large federally-funded energy-efficiency program in the US which targets low-income families and provides full subsidies for improving the conditions of the HVAC (heating, ventilation and air conditioning) systems of their homes. This is an ideal setting for demonstrating the properties of the method proposed in this paper for a few reasons. First, WAP allows for a data-rich environment: I have access to data from over 34 thousand homes served by WAP from 2006 to 2016 in the state of Illinois. Detailed information about these homes are available, including energy billing, housing structure, local weather, and demographic variables.[25]

[25] Without considering interactions or transformations, I have access to 29 variables, described in detail in Appendix F. Summary statistics of the outcome of interest (monthly natural gas usage) and available covariates can be found in Appendix E.

Second, these homes were served by the program at different points in time (i.e., constituting staggered adoption). Finally, this is a setting in which significant treatment effect heterogeneity is expected. For example, effects may vary depending on year of treatment, due to differences in program implementation guidelines. Depreciation of the HVAC systems may also play a role. Importantly, effects are expected to significantly vary across homes, which may inherently need different types of upgrades.

For this application, the outcome of interest is natural gas usage, measured in MMBtu. The main objective is to recover heterogeneity of effects across homes that received different levels of WAP spending on diverse measures. I also investigate heterogeneity of savings across housing structure and demographics. For that purpose, I estimate conditional average treatment effects on the treated (CATT). The CATT estimates are then used for measure-specific cost-benefit analyses. XGBoost is used for the first (predictive) step of the method.[26] To train the (counterfactual) model, I restrict the sample to all (actual) pre-treatment observations. Further I restrict the sample to observations within a window of 2 years before and after treatment. This helps with the argument that Assumption 4 (stability of counterfactual function) is likely to hold, and implies that I focus on near-term estimates of the effects of the program. With the tests proposed in section 3.1.2, in Appendix F, I show that prediction errors are unlikely to be correlated with the available covariates of interest.

5.1 Estimates of Heterogeneous Treatment Effects

To assess heterogeneity of program savings, I estimate CATT according to step three from my proposed ML approach. Specifically, after obtaining home-by-month treatment effects, I run the following linear regression to decompose them:

$$\hat{b}_{i,t} = \alpha_0 + \sum_{k=1}^{K} \beta_k C_{i,t}^k + \sum_{g=1}^{G} \gamma_g X_{i,t}^g + u_{i,t} \text{ , for all } D_{i,t} = 1 \text{ ,} \tag{8}$$

where $\hat{b}_{i,t}$ are natural gas savings (MMBtu) for home i in the post-treatment ($D_{i,t} = 1$)

[26] In Appendix F, I show that XGBoost achieved high out-of-sample (cross-validated) prediction accuracy for this real data application.

months t; α_0 is a constant; $X_{i,t}^g$ includes the following covariates: housing structure (air sealing, blower door reading, attic R-value, floor area, number of stories, heating unit size, and vintage); demographics (household income, householder age, and family size); natural gas and electricity prices; and weather controls (average minimum temperature, average maximum temperature, and average precipitation). $C_{i,t}^k$ are categories of program spending: air conditioning, air sealing, attic, baseload, doors, foundation, furnace, health and safety, wall insulation, water heater, windows, and other incidentals. Variables are flexibly included via binning. Bins can vary in size, depending on the distribution of the variable considered.

I compare the machine learning estimates with those from a fully interacted two-way fixed effects model where I regress natural gas consumption on home by calendar month FE, plus month of sample by county FE, in addition to covariates interacted with the binary treatment indicator. For this TWFE estimator, the coefficients of interest are those associated with the interactions between treatment and covariates (especially related to program spending).

Figures 3 and 4 present estimates of heterogeneous treatment effects for selected upgrades or home characteristics. I focus on covariates that are expected to be closely related to energy consumption. The graphs should be interpreted as follows: the vertical axes represent natural gas savings (MMBtu) attributed to WAP, while the horizontal axes represent bins of amount spent on upgrades or other relevant home characteristics. To avoid collinearity, for each variable it was necessary to drop one of the bins, to serve as the omitted comparison group the estimating equation. For the spending categories, I drop the first bin of zero amount spent. For all other cases, I drop the bin which includes the median value along a given dimension. The presented coefficients should be interpreted as heterogeneity in energy savings, compared to the omitted bin. Blue triangles represent coefficients according the ML estimator, while the red squares are those from TWFE.

First, comparing ML versus TWFE coefficients, it can be noted that they generally trend in the same directions, and reveal strikingly similar patterns of heterogeneity. However, ML estimates are more precise. Furthermore, there are some notable discrepancies

between the estimates. Focusing on Wall Insulation and Attics, for example, TWFE suggest stronger treatment effects compared to ML. That may be attributed to TWFE not accurately capturing the temporal variation of effects. As simulation results from Section 4 reveal, TWFE coefficients will be overestimated in case the true underlying effects are stronger in the months right after treatment. That scenario is consistent with depreciation of the upgrades performed by WAP.

Focusing on interpreting the preferred ML specification, the graphs reveal several interesting patterns of heterogeneity for this context, which had not been previously documented in the literature. When looking at furnaces, for example, it can be noted that spending below $1,500 is associated with an *increase* in energy usage. On the other hand, significant energy savings are achieved with furnace spending above $1,800. Lower levels of furnace spending correspond to repair and re-tuning, which may be associated with rebound effects (residents using their furnaces more often), without substantial improvement to the efficiency of the furnace. However, high levels of furnace spending correspond to installing new (likely more efficient) furnaces, thus leading to significant reduction in energy consumption.[27] That is an intuitive result but the magnitudes or the importance of the savings from replacing furnaces should not be understated. In section 5.2, I provide more insight about the cost-effectiveness of furnace replacements versus repairs.

Graphs labelled as Attic, Wall Insulation, and Foundation collectively represent the majority of insulation measures performed by the program. As expected, those reveal that insulation is crucial for energy savings in the context of WAP. Virtually any level of insulation spending is associated with some energy savings. Further, the relationship between savings and spending on insulation seems to be mostly linear. Only high levels of spending (above $1,200) on windows are significantly associated with some energy savings.

I also show that homes with a larger pre-treatment heating unit achieve better

[27] Appendix E presents the histograms for each of the WAP spending categories. It can be noted, for example, that the distribution is bimodal for furnaces, thus suggesting a separation between simple repair/re-tuning versus complete replacements. This is corroborated by assessing more specific descriptions of measures performed in each home, available in the raw Program administrative data.

energy savings. Those large units may therefore have been replaced with smaller ones (which use less energy). Otherwise, the units may have been replaced by newer models that are more efficient regardless of size. That is consistent with the results from furnace spending.

In terms of demographics, the machine learning estimates suggest U-shaped relationships between energy consumption and family size, as well as between energy energy consumption and householder age. Compared to the median, both younger and older householders, as well as smaller and bigger families consume more energy after treatment. The differences along those dimensions are small, nevertheless significant. Coefficients on householder age from TWFE differ substantially from those from machine learning. However, potential sources of bias along those dimensions are unclear.

[FIGURE 3 HERE]

[FIGURE 4 HERE]

5.2 Upgrade-Specific Cost-Benefit Analyses

In this section, I investigate if each of the categories of WAP investments are cost-effective. Measure-specific costs were obtained from administrative data. They incorporate both labor and materials costs. I assume that benefits accrue through reduced energy savings only, according to the parameters estimated in the above section.[28] I focus on the measures that were associated with significant energy savings.

For each measure and each bin of spending, I compute the monetized benefits of reduced natural gas usage. I take into account social marginal benefits, incorporating the social costs of carbon following the procedure as described in Davis and Muehlegger (2010). The average citygate natural gas prices in Illinois from 2007-2016 represent marginal private costs, to which I add the social costs of carbon of $40 per ton. Emissions

[28] WAP may also be associated with indoor air quality, and health benefits, for example. The investigation of those benefits is left for future research.

factors for natural gas were obtained from EPA (1998). The resulting price is assumed for the first post-treatment month, after which escalation is applied based on indices from Rushing, Kneifel, and Lippiatt (2012).

Different measures are assumed to have different lifespans. Baseline scenarios follow lifespan recommendations from official WAP documentation: 25 years for insulation measures; 20 years for furnaces; 15 years for windows. Measures are assumed to fully depreciate after those lifespans. However, there is uncertainty regarding those lifespans, and recent engineering literature suggests that they could be longer (Kono et al., 2016). Therefore, I also consider the following alternative lifespans: 50 years for insulation; 30 years for furnaces and windows. Finally, to obtain the present value of benefits, I use a discount rate of 3%, which is the recommended rate for evaluation of several governmental programs, including WAP (Rushing, Kneifel, and Lippiatt, 2012).

I subtract monetized benefits from per-measure costs to obtain net benefits for all the bins of spending.[29] Results are presented in Figure 5. First, it can be noted that only insulation measures, especially for attics, are associated with positive net benefits. Attic spending exhibits a clear pattern of diminishing returns. Further, net benefits are sensitive to lifespan assumptions. Comparing both lifespan scenarios, the difference in net benefits can be up to $3,000 for attics, for example. Foundation and wall insulation are at the margin of cost-effectiveness with baseline assumptions. With longer lifespans, those measure are therefore associated with positive net benefits.

Furnace and windows are generally associated with negative net benefits. The bimodal distribution for furnace is again clear in these cost-benefit analyses, suggesting that expensive furnace repairs ($600 - $1,800) are less cost-effective than full furnace replacements (above $1,800). In this context, negative net benefits do not necessarily imply that some measures should be performed. It is important to note that WAP measures may be complementary. For example, better wall insulation can enhance the benefits from a more efficient furnace. Analyses of complex interactions between measures are left for future work.

[29] I use average costs within each bin.

The methods and results presented in this paper complement the analyses in Christensen et al. (2021). That paper provides insight about the mechanisms that can explain a wedge between *ex-ante* projected and *ex-post* realized energy savings from WAP. Results suggest that biases in projected savings are especially associated with systematic engineering modelling errors and workmanship, while changes in consumer behavior (rebound effects) are not significant in this setting.

[FIGURE 5 HERE]

6 Conclusions

I introduce a novel method to estimate heterogeneous treatment effects for event studies with staggered adoption. I contribute to a growing literature that proposes alternatives to the standard TWFE in these settings. The proposed method employs highly flexible machine learning algorithms to predict counterfactuals, which in turn are used to estimate treatment effects. Within this framework, I propose tests to assess the validity of the assumptions required to identify certain causal parameters of interest. Further, I emphasize the role of cross-validation to assess the performance of the model for counterfactual predictions, and to account for that model's potential errors when performing inference. I perform my analyses within a data-rich environment, which allows a deep exploration of the several dimensions of heterogeneity.

With simulations using publicly-available air pollution data from Spain, I test the performance of the proposed machine learning method, contrasting it with standard two-way fixed effects regression and with imputation approaches that model counterfactuals through OLS. I show that, consistent with prior literature, TWFE can be near-term biased in cases where treatment effects are dynamic (time-varying). Conversely, the ML and OLS imputation approaches are shown to be unbiased. Further, my ML based proposal is more efficient than imputation via standard OLS. Other advantages of the ML approach, in particular, are: it allows the researcher to be agnostic about the specification

of the model for counterfactuals; provides a straightforward framework for assessing the validity of identifying assumptions; and is less subject to loss of observations (as one may not need to include unit fixed effects in the ML specifications).

I conclude with an application of the ML approach to real data from the Weatherization Assistance Program. I am able to identify substantial heterogeneity of program effects, which had not been empirically documented in the literature. For example, I find that even though insulation measures are among the most important drivers of energy savings in this program, the cost-effectiveness of these measures is sensitive to assumptions regarding their lifespans. I also find evidence that furnace replacements are more cost-effective than particularly expensive furnace repairs/re-tuning. Since the ML method allows estimation of fine-scale heterogeneity, it may be useful to aid in an exercise to identify high-return homes, to which funds may be targeted more cost-effectively. I also reiterate that the approach proposed in this paper is not only applicable to research in energy and environmental economics. Rather, the method can be considered for recovering heterogeneity in event studies within data-rich environments, as long as the identifying assumptions hold.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". *Journal of the American Statistical Association* 105(490), pp. 493–505.

Abrell, Jan, Mirjam Kosch, and Sebastian Rausch (2022). "How effective is carbon pricing?—A machine learning approach to policy evaluation". *Journal of Environmental Economics and Management* 112, p. 102589. URL: https://www.sciencedirect.com/science/article/pii/S0095069621001339.

Agencia Estatal de Meteorologia (2020). "Climatologías Diarias". *AEMET OpenData*. URL: https://opendata.aemet.es/centrodedescargas/inicio.

Allcott, Hunt and Michael Greenstone (2017). "Measuring the Welfare Effects of Residential Energy Efficiency Programs". *NBER Working Paper*(23386).

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), pp. 236–76.

Angrist, J. D. and J. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Arlot, Sylvain and Alain Celisse (2010). "A survey of cross-validation procedures for model selection". *Statistics Surveys* 4, pp. 40–79. URL: https://doi.org/10.1214/09-SS054.

Athey, Susan (2019). "21. The Impact of Machine Learning on Economics". *The Economics of Artificial Intelligence: An Agenda*. Ed. by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. University of Chicago Press, pp. 507–552. URL: https://doi.org/10.7208/9780226613475-023.

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu (2019). "Ensemble Methods for Causal Effects in Panel Data Settings". *arXiv Working Paper*. URL: https://arxiv.org/abs/1903.10079.

Athey, Susan and Guido Imbens (2016). "Recursive partitioning for heterogeneous causal effects". *Proceedings of the National Academy of Sciences* 113(27), pp. 7353–7360. URL: https://www.pnas.org/content/113/27/7353.

Athey, Susan and Guido W. Imbens (2019). "Machine Learning Methods That Economists Should Know About". *Annual Review of Economics* 11(1), pp. 685–725. eprint: https://doi.org/10.1146/annurev-economics-080217-053433. URL: https://doi.org/10.1146/annurev-economics-080217-053433.

Athey, Susan and Guido W. Imbens (2022). "Design-based analysis in Difference-In-Differences settings with staggered adoption". *Journal of Econometrics* 226(1). Annals Issue in Honor of Gary Chamberlain, pp. 62–79. URL: https://www.sciencedirect.com/science/article/pii/S0304407621000488.

Baker, Andrew, David F. Larcker, and Charles C. Y. Wang (2022). "How Much Should We Trust Staggered Difference-In-Differences Estimates?" *Journal of Financial Economics*. (Forthcoming). URL: http://dx.doi.org/10.2139/ssrn.3794018.

Balzer, Laura B., Maya L. Petersen, and Mark J. van der Laan (2016). "Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching". *Statistics in Medicine* 35(21), pp. 3717–3732.

Biau, Gérard and Benoît Cadre (2021). "Optimization by Gradient Boosting". *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan*. Ed. by Abdelaati Daouia and Anne Ruiz-Gazen. Springer International Publishing, pp. 23–44. URL: https://doi.org/10.1007/978-3-030-73249-3_2.

Borusyak, Kirill and Xavier Jaravel (2017). "Revisiting Event Study Designs". *SSRN Working Paper*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2826228.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess (2021). "Revisiting Event Study Designs: Robust and Efficient Estimation". *arXiv Working Paper 2108.12419*. URL: https://arxiv.org/abs/2108.12419.

Burlig, Fiona, Christopher Knittel, David Rapson, Mar Reguant, and Catherine Wolfram (2020). "Machine Learning from Schools about Energy Efficiency". *Journal of the Association of Environmental and Resource Economists* 7(6), pp. 1181–1217. eprint: https://doi.org/10.1086/710606. URL: https://doi.org/10.1086/710606.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro H. C. Sant'Anna (2021). "Difference-in-Differences with a Continuous Treatment". *arXiv:2107.02637*.

Callaway, Brantly and Pedro H.C. Sant'Anna (2021). "Difference-in-Differences with Multiple Time Periods". *Journal of Econometrics* 225(2), pp. 200–230. URL: https://www.sciencedirect.com/science/article/pii/S0304407620303948.

Chaisemartin, Clément de and Xavier D'Haultfoeuille (2021). "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey". *arXiv Working Paper*. URL: https://arxiv.org/abs/2112.04565.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". *arXiv:1603.02754*.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey (2017). "Double/Debiased/Neyman Machine Learning of Treatment Effects". *American Economic Review* 107(5), pp. 261–65. URL: http://www.aeaweb.org/articles?id=10.1257/aer.p20171038.

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo (2018). "The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages". *Econometrica* 86(6), pp. 1911–1938.

Christensen, Peter, Paul Francisco, Erica Myers, and Mateus Souza (2021). "Decomposing the Wedge between Projected and Realized Returns in Energy Efficiency Programs". *The Review of Economics and Statistics*. (Forthcoming), pp. 1–46. URL: https://doi.org/10.1162/rest%5C_a%5C_01087.

Coble, Keith H, Ashok K Mishra, Shannon Ferrell, and Terry Griffin (2018). "Big Data in Agriculture: A Challenge for the Future". *Applied Economic Perspectives and Policy* 40(1), pp. 79–96. URL: https://doi.org/10.1093/aepp/ppx056.

Davis, Lucas W. and Erich Muehlegger (2010). "Do Americans consume too little natural gas? An empirical test of marginal cost pricing". *The RAND Journal of Economics* 41(4), pp. 791–810. URL: http://www.jstor.org/stable/25746054.

de Chaisemartin, Clément and Xavier D'Haultfoeuille (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". *American Economic Review* 110(9), pp. 2964–96. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20181169.

ESIOS (2020). "Generation and Consumption Data". *Red Elétrica de España; Sistema de Informacíon del Operador del Sistema*. URL: https://www.esios.ree.es/en/generation-and-consumption.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram (2018). "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program". *The Quarterly Journal of Economics* 133(3), pp. 1597–1644.

Gardner, John (2021). "Two-stage differences in differences". *Working Paper*. URL: https://jrgcmu.github.io/2sdd_current.pdf.

Ghanem, Dalia and Aaron Smith (2021). "What Are the Benefits of High-Frequency Data for Fixed Effects Panel Models?" *Journal of the Association of Environmental and Resource Economists* 8(2), pp. 199–234. URL: https://doi.org/10.1086/710968.

Ghoddusi, Hamed, Germán G. Creamer, and Nima Rafizadeh (2019). "Machine learning in energy economics and finance: A review". *Energy Economics* 81, pp. 709–727. URL: https://www.sciencedirect.com/science/article/pii/S0140988319301513.

Gillingham, Kenneth, David Rapson, and Gernot Wagner (2016). "The Rebound Effect and Energy Efficiency Policy". *Review of Environmental Economics and Policy* 10(1), pp. 68–88. URL: https://doi.org/10.1093/reep/rev017.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". *Journal of Econometrics* 225(2), pp. 254–277. URL: https://www.sciencedirect.com/science/article/pii/S0304407621001445.

Hyndman, Rob J and George Athanasopoulos (2018). *Forecasting: principles and practice*. OTexts. Chap. 5.10 - Time series cross-validation. URL: https://otexts.com/fpp3/tscv.html.

Imai, Kosuke and In Song Kim (2020). "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data". *Political Analysis*. URL: https://doi.org/10.1017/pan.2020.33.

Institutio Nacional de Estadística (2020). "Demografía y población". URL: https://www.ine.es/index.htm.

Jin, Xiaolong, Benjamin W. Wah, Xueqi Cheng, and Yuanzhuo Wang (2015). "Significance and Challenges of Big Data Research". *Big Data Research* 2(2), pp. 59–64. URL: http://www.sciencedirect.com/science/article/pii/S2214579615000076.

Kono, Jun, Yutaka Goto, York Ostermeyer, Rolf Frischknecht, and Holger Wallbaum (2016). "Factors for Eco-Efficiency Improvement of Thermal Insulation Materials". *Key Engineering Materials* 678, pp. 1–13.

Kropko, Jonathan and Robert Kubinec (2020). "Interpretation and identification of within-unit and cross-sectional variation in panel data models". *Plos One*. URL: https://doi.org/10.1371/journal.pone.0231349.

Liu, Licheng, Ye Wang, and Yiqing Xu (2021). "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data". *arXiv Working Paper*. URL: https://arxiv.org/abs/2107.00856.

Marcus, Michelle and Pedro H. C. Sant'Anna (2021). "The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics". *Journal of the Association of Environmental and Resource Economists* 8(2), pp. 235–275. URL: https://doi.org/10.1086/711509.

Miller, Steve (2020). "Causal forest estimation of heterogeneous and time-varying environmental policy effects". *Journal of Environmental Economics and Management* 103, p. 102337. URL: https://www.sciencedirect.com/science/article/pii/S0095069620300607.

Ministerio para la Transición Ecológica y el Reto Demográfico (2020). "Datos Calidad del Aire 2001-2019". URL: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/evaluacion-datos/datos/Datos_2001_2019.aspx.

Ministerio para la Transición Ecológica y el Reto Demográfico (2021). "Incendios forestales, en datos, estadísticas y cifras". URL: https://www.epdata.es/datos/incendios-forestales-datos-estadisticas-cifras/267?accion=2.

Mundlak, Yair (1978). "On the Pooling of Time Series and Cross Section Data". *Econometrica* 46(1), pp. 69–85. URL: http://www.jstor.org/stable/1913646.

Neyman, J. (1923). "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9". *Roczniki Nauk Rolniczych Tom X [Polish]; translated in Statistical Science (1990)* 5, pp. 465–480.

Polley, Eric, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan (2018). "SuperLearner: Super Learner Prediction". *The Comprehensive R Archive Network (CRAN)*. URL: https://CRAN.R-project.org/package=SuperLearner.

Prest, Brian C. (2020). "Peaking Interest: How Awareness Drives the Effectiveness of Time-of-Use Electricity Pricing". *Journal of the Association of Environmental and Resource Economists* 7(1), pp. 103–143. URL: https://doi.org/10.1086/705798.

Robins, James (1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Mathematical Modelling* 7(9), pp. 1393–1512.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". *Journal of Educational Psychology* 66(5), pp. 668–701.

Rushing, Amy S., Joshua D. Kneifel, and Barbara C. Lippiatt (2012). "Energy Price Indices and Discount Factors for Life-Cycle Cost Analysis – 2012: Annual Supplement to NIST Handbook 135 and NBS Special Publication 709". *NIST Interagency/Internal Report (NISTIR)* 15(n29). URL: https://nvlpubs.nist.gov/nistpubs/ir/2012/NIST.IR.85-3273-27.pdf.

Solon, Gary, Steven J Haider, and Jeffrey M Wooldridge (2015). "What are we weighting for?" *Journal of Human Resources* 50(2), pp. 301–316.

Storm, Hugo, Kathy Baylis, and Thomas Heckelei (2019). "Machine learning in agricultural and applied economics". *European Review of Agricultural Economics*. URL: https://doi.org/10.1093/erae/jbz033.

Strezhnev, Anton (2018). "Semiparametric weighting estimators for multi-period difference-in-differences designs". *Working Paper*. URL: https://www.antonstrezhnev.com/research/.

Sun, Liyang and Sarah Abraham (2021). "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects". *Journal of Econometrics* 225(2). Themed Issue: Treatment Effect 1, pp. 175–199. URL: https://www.sciencedirect.com/science/article/pii/S030440762030378X.

US Environmental Protection Agency (1998). "AP 42, Fifth Edition Compilation of Air Pollutant Emissions Factors, Volume 1: Stationary Point and Area Sources". *Technical Report*. URL: https://www3.epa.gov/ttn/chief/ap42/ch01/final/c01s04.pdf.

van der Laan, Mark and Daniel Rubin (2006). "Targeted Maximum Likelihood Learning". *The International Journal of Biostatistics* 2(1). URL: https://doi.org/10.2202/1557-4679.1043.

Varian, Hal R. (2014). "Big Data: New Tricks for Econometrics". *Journal of Economic Perspectives* 28(2), pp. 3–28. URL: https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3.

Wager, Stefan and Susan Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* 113(523), pp. 1228–1242.

Weersink, Alfons, Evan Fraser, David Pannell, Emily Duncan, and Sarah Rotz (2018). "Opportunities and Challenges for Big Data in Agricultural and Environmental Analysis". *Annual Review of Resource Economics* 10(1), pp. 19–37. URL: https://doi.org/10.1146/annurev-resource-100516-053654.

Wooldridge, Jeff (2021). "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators". *SSRN Working Paper 3906345*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3906345.

Yu, Zhuo and Mark J. van der Laan (2002). "Construction of Counterfactuals and the G-computation Formula". *U.C. Berkeley Division of Biostatistics Working Paper Series* Working Paper 122. URL: https://biostats.bepress.com/ucbbiostat/paper122.

Zivin, Joshua G. and Kevin Novan (2016). "Upgrading Efficiency and Behavior: Electricity Savings from Residential Weatherization Programs". *The Energy Journal* 37(4).

Tables

Table 1: Simulation Results – Performance of Estimators for ATT Under Dynamic Treatment Effects

	(1) Simulated (benchmark)	(2) Machine Learning	(3) OLS TWFE	(4) OLS TWFE (saturated)	(5) OLS Imputation	(6) OLS Imputation (saturated)
Panel A: Treatment Effects Decreasing Over Time						
$\widehat{ATT}$	-3.1840	-3.1350	-4.9006	-4.5487	-3.0367	-2.7432
Standard Errors		(0.1281)	(0.2891)	(0.3457)	(0.2548)	(0.2118)
Bootstrapped Standard Errors		[0.2272]				
Observations	170,484	170,484	170,484	154,999	170,133	128,718
Panel B: Treatment Effects Increasing Over Time						
$\widehat{ATT}$	-2.2199	-2.1709	-0.9287	-0.4429	-2.0691	-1.6989
Standard Errors		(0.1244)	(0.2869)	(0.3334)	(0.2563)	(0.2153)
Bootstrapped Standard Errors		[0.2199]				
Observations	170,484	170,484	170,484	154,999	170,133	128,718
Station FE		NA	Yes	No	Yes	No
Day of sample FE		NA	Yes	No	Yes	No
Station × Month FE		NA	No	Yes	No	Yes
Day of sample × Province FE		NA	No	Yes	No	Yes
Additional controls		NA	No	Yes	No	Yes

Notes: This table presents the performance of alternative methods for estimating ATT under dynamic treatment effects. The outcome variable is PM_{10} particulate matter concentrations, measured in $\mu g/m^3$. The simulations impose a reduction in PM_{10} for the post-treatment sample. The ATT aims to recover the full post-treatment sample average of that reduction. Panel A is for results with treatment effects that decrease in magnitude over time, while Panel B is for results with treatment effects that increase over time. Column (1) presents the "true" effect which serves as the benchmark. Column (2) presents results from my ML approach. Columns (3) and (4) are for standard TWFE approaches. Columns (5) and (6) are for an imputation approach that uses OLS for prediction of untreated counterfactuals. Standard errors from standard TWFE are clustered by air quality station. Standard errors for the ML approach are also clustered by station and are further adjusted using cross-validated residuals from the predictive step, according to equation (6). Conservative bootstrapped standard errors (200 iterations) for the ML approach are presented in square brackets. Standard errors for the OLS imputation approach are adjusted, but using in-sample residuals from the predictive step. As described in section 3.4, this adjustment also takes into account deviations from an average effect ($\hat{\bar{b}}_{i,t}$), which I calculate as averages for each month of sample.

Table 2: Simulation Results – Performance of Estimators for Heterogeneous Treatment Effects Over Time

	(1) Simulated (benchmark)	(2) Machine Learning	(3) OLS TWFE	(4) OLS TWFE (saturated)	(5) OLS Imputation	(6) OLS Imputation (saturated)
$\widehat{ATT}(1)$: Semester 1	-4.2808	-4.2578	-4.5397	-4.0435	-4.2057	-3.5697
Standard Errors		(0.1568)	(0.2967)	(0.3733)	(0.2386)	(0.3176)
Bootstrapped Standard Errors		[0.2258]				
$\widehat{ATT}(2)$: Semester 2	-3.1807	-3.2693	-3.4408	-2.9358	-3.0351	-2.8437
		(0.1814)	(0.4054)	(0.4179)	(0.2969)	(0.3106)
		[0.2572]				
$\widehat{ATT}(3)$: Semester 3	-2.1932	-2.0567	-2.3363	-1.7575	-1.9510	-1.6892
		(0.2619)	(0.5098)	(0.5406)	(0.4209)	(0.4430)
		[0.3369]				
$\widehat{ATT}(4)$: Semester 4	-1.1405	-0.8112	-1.3540	-0.2796	-0.8908	-0.8406
		(0.3677)	(0.5721)	(0.5758)	(0.5169)	(0.4410)
		[0.4187]				
Observations	170,484	170,484	170,484	154,999	170,133	128,718
Station FE		NA	Yes	No	Yes	No
Day of sample FE		NA	Yes	No	Yes	No
Station × Month FE		NA	No	Yes	No	Yes
Day of sample × Province FE		NA	No	Yes	No	Yes
Additional controls		NA	No	Yes	No	Yes

Notes: This table presents the performance of alternative methods for estimating ATT(r). That is, for recovering a different effect for each semester after treatment. The outcome variable is PM_{10} particulate matter concentrations, measured in $\mu g/m^3$. The simulation imposes an effect that becomes weaker in magnitude over time, as shown in column (1) of the "true" benchmark effects. Column (2) presents results from my ML approach. Columns (3) and (4) are for standard TWFE approaches. Columns (5) and (6) are for an imputation approach that uses OLS for prediction of untreated counterfactuals. Standard errors from standard TWFE are clustered by air quality station. Standard errors for the ML approach are also clustered by station and are further adjusted using cross-validated residuals from the predictive step, according to equation (6). Conservative bootstrapped standard errors (200 iterations) for the ML approach are presented in square brackets. Standard errors for the OLS imputation approach are adjusted, but using in-sample residuals from the predictive step. As described in section 3.4, this adjustment also takes into account deviations from an average effect ($\hat{\bar{b}}_{i,t}$), which I calculate as averages for each month of sample.

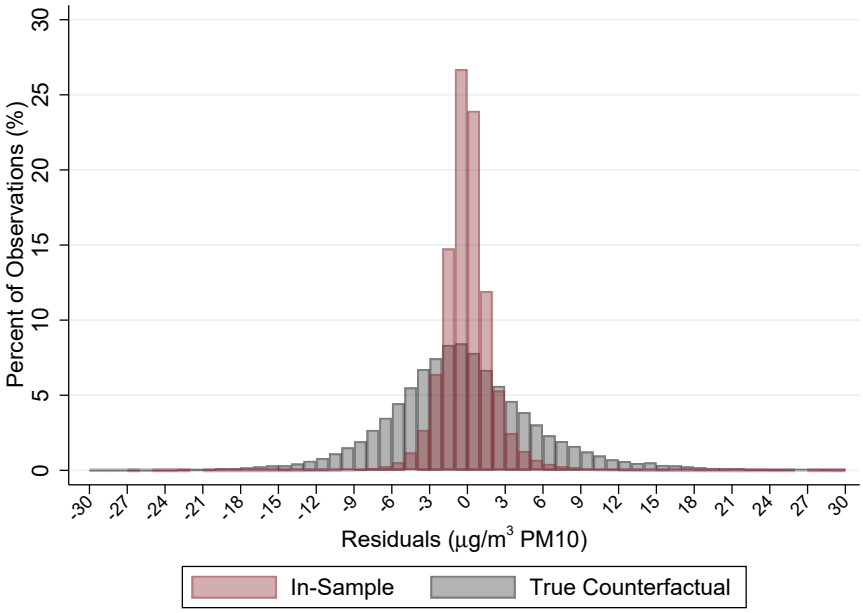
Table 3: Simulation Results – Performance of Estimators for CATT

	(1) Simulated (benchmark)	(2) Machine Learning	(3) OLS TWFE	(4) OLS TWFE (saturated)	(5) OLS Imputation	(6) OLS Imputation (saturated)
$\widehat{CATT}$: Altitude $\leq$ 35m	-7.0897	-6.9100	-8.3251	-7.9242	-6.6679	-6.4911
Standard Errors		(0.2885)	(0.4644)	(0.4796)	(0.4471)	(0.4048)
Bootstrapped Standard Errors		[0.4069]				
$\widehat{CATT}$: 35m $<$ Altitude $\leq$ 150m	-4.1586	-4.0815	-5.5698	-5.4877	-4.6664	-4.2504
		(0.1982)	(0.3555)	(0.4692)	(0.5500)	(0.2942)
		[0.2891]				
$\widehat{CATT}$: 150m $<$ Altitude $\leq$ 500m	-1.2035	-1.1198	-2.8135	-2.7120	-1.0951	-0.4605
		(0.2662)	(0.4679)	(0.6323)	(0.4719)	(0.5502)
		[0.4055]				
$\widehat{CATT}$: Altitude $>$ 500m	0.0000	-0.2662	-1.3353	-0.2348	0.8703	1.2154
		(0.3053)	(0.4479)	(0.8139)	(0.4615)	(0.4980)
		[0.5459]				
Observations	170,484	170,484	170,484	154,999	170,133	128,718
Station FE		NA	Yes	No	Yes	No
Day of sample FE		NA	Yes	No	Yes	No
Station $\times$ Month FE		NA	No	Yes	No	Yes
Day of sample $\times$ Province FE		NA	No	Yes	No	Yes
Additional controls		NA	No	Yes	No	Yes

Notes: This table presents the performance of alternative methods for estimating conditional average treatment effects on the treated (CATT). That is, for recovering a different effect for each group that identifies the altitude of an air quality station. The outcome variable is PM_{10} particulate matter concentrations, measured in $\mu g/m^3$. The simulation imposes an effect that becomes weaker in magnitude over time, and that varies with altitude. Column (1) shows the "true" benchmark effects for each group. Column (2) presents results from my ML approach. Columns (3) and (4) are for standard TWFE approaches. Columns (5) and (6) are for an imputation approach that uses OLS for prediction of untreated counterfactuals. Standard errors from standard TWFE are clustered by air quality station. Standard errors for the ML approach are also clustered by station and are further adjusted using cross-validated residuals from the predictive step, according to equation (6). Conservative bootstrapped standard errors (200 iterations) for the ML approach are presented in square brackets. Standard errors for the OLS imputation approach are adjusted, but using in-sample residuals from the predictive step. As described in section 3.4, this adjustment also takes into account deviations from an average effect ($\hat{\tilde{b}}_{i,t}$), which I calculate as averages for each of the four groups defined based on altitude.

Figures

Panel A: In-Sample Residuals



Panel B: Cross-Validated Residuals

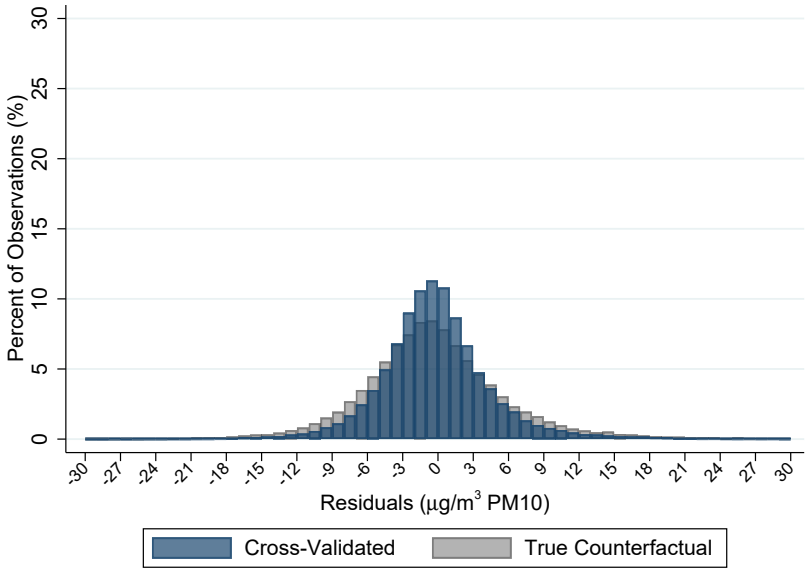


Figure 1: Histograms for In-Sample, Cross-Validated, and True Counterfactual Residuals from the ML Approach

Notes: This Figure presents histograms of residuals for the first step of my ML approach, applied within the simulation setting. These are residuals according to the best-performing XGBoost configuration for predicting PM_{10} particulate matter concentrations (measured in $\mu g/m^3$). Panel A compares in-sample (in red) and true counterfactual (in gray) residuals. Panel B compares cross-validated (in blue) and true counterfactual (in gray) residuals.

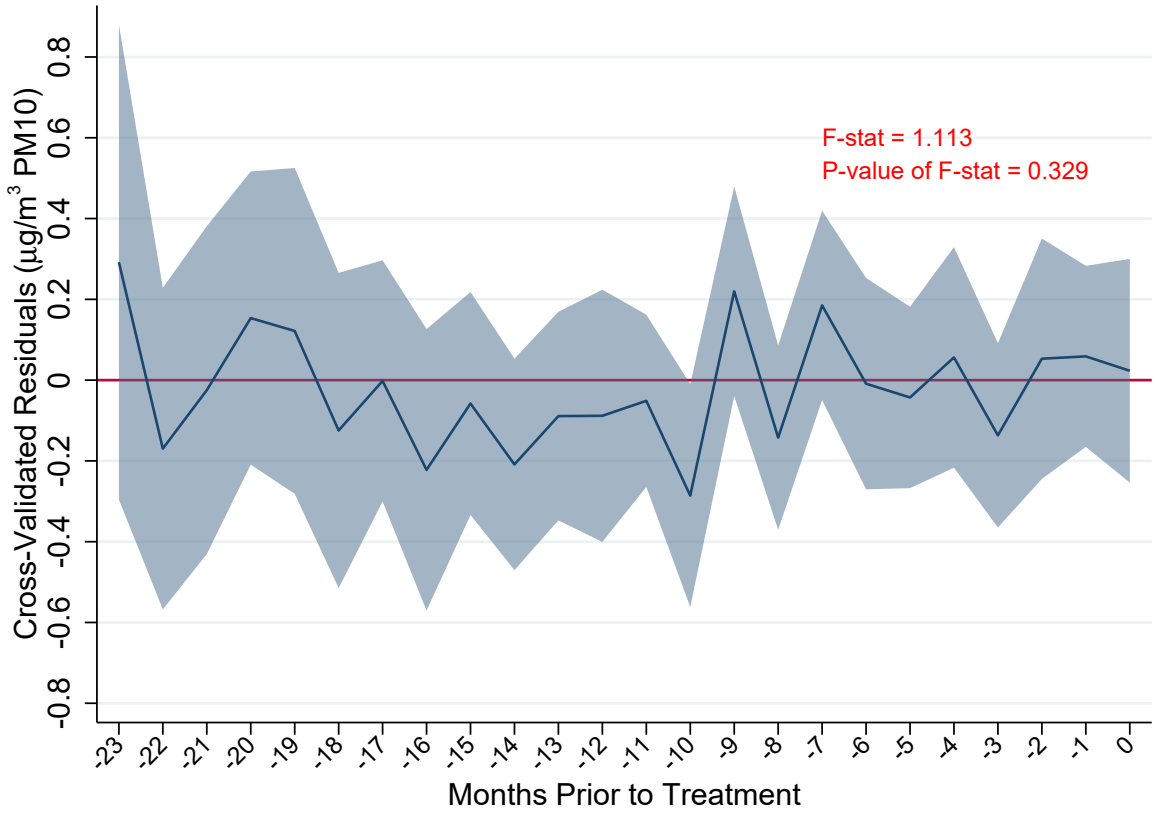


Figure 2: Assessing Anticipatory Effects and the Stability of the Counterfactual Function

Notes: This Figure plots coefficient estimates and 95% confidence intervals from a regression of cross-validated residuals on indicators for time relative to treatment (equation 3). This is for testing Assumption 2 (no anticipatory effects), and Assumption 4 (stability of the counterfactual function). This is analogous to "pre-trends" tests in traditional difference-in-differences settings. The top right corner of the Figure shows the resulting F-statistic and associated P-value for a test of joint significance of the coefficients.

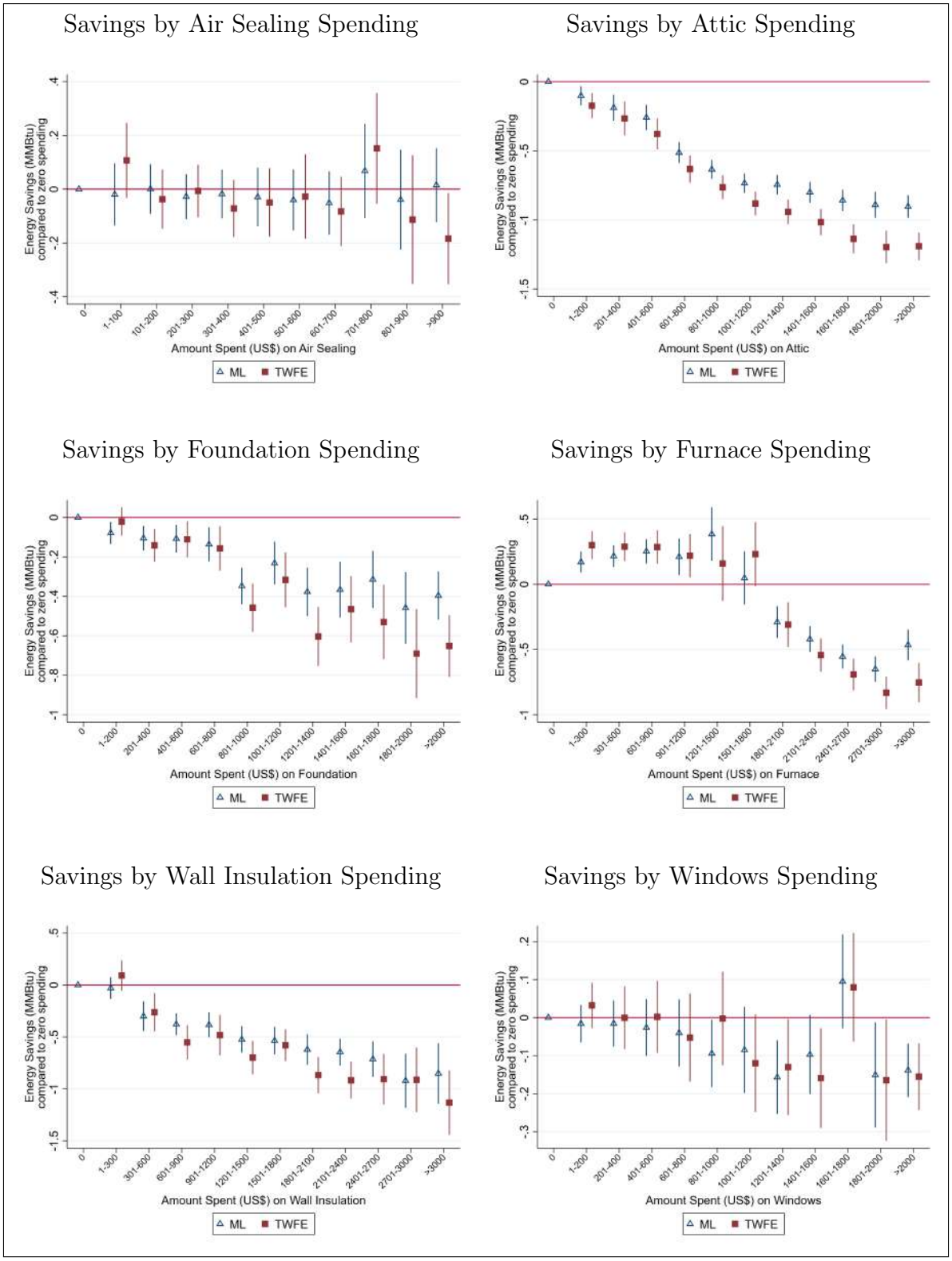


Figure 3: ML Heterogeneous Treatment Effect Estimates for Program Spending

Notes: The figures above present machine learning estimates of heterogeneous treatment effects for selected WAP categories of spending. Negative coefficients should be interpreted as percent energy savings attributed to WAP treatment. ML standard errors were bootstrapped (200 iterations). For two-way fixed effects, standard errors are clustered by household.

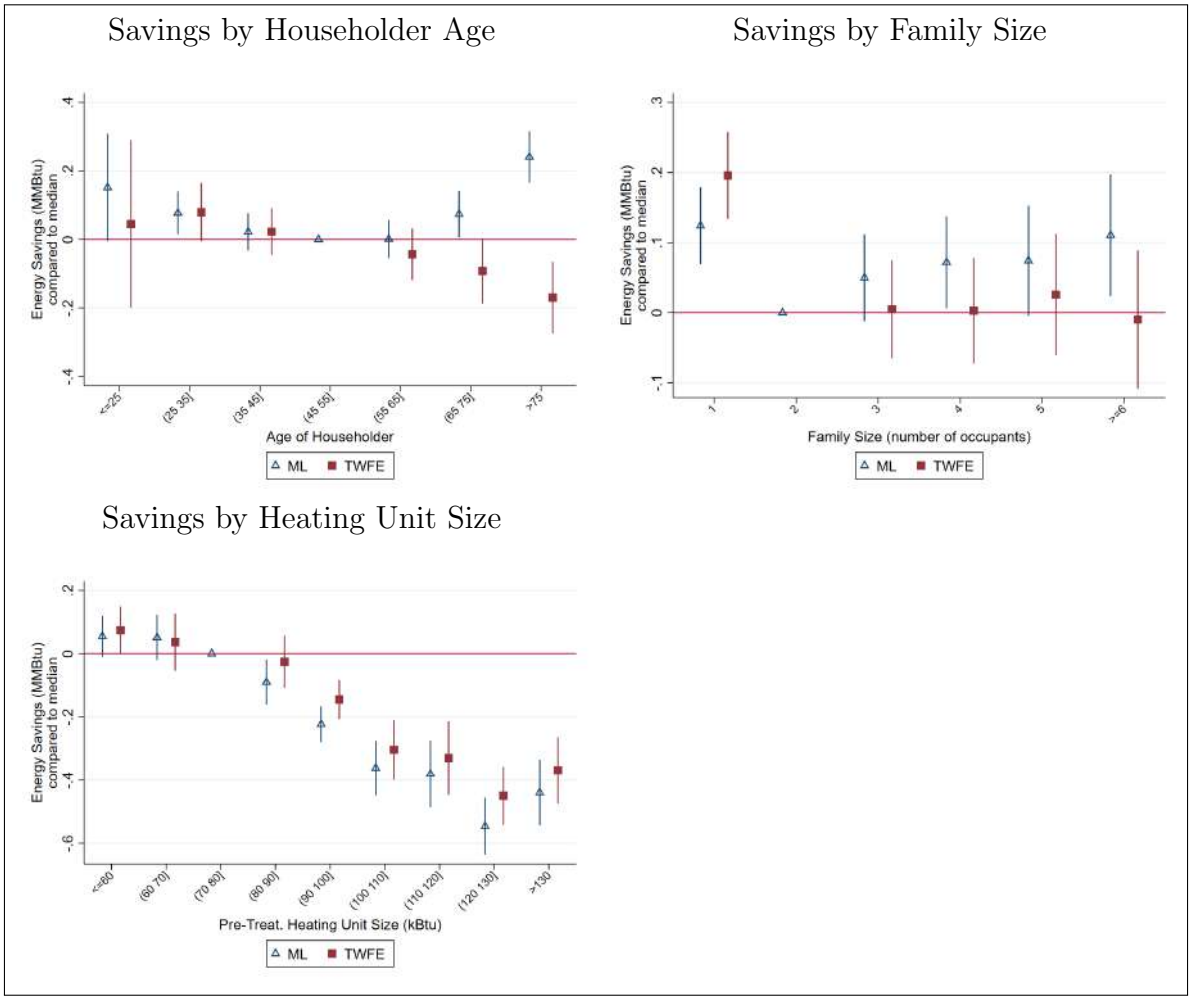


Figure 4: ML Heterogeneous Treatment Effect Estimates for Selected Covariates

Notes: The figures above present machine learning estimates of heterogeneous treatment effects for selected covariates. Negative coefficients should be interpreted as percent energy savings attributed to WAP treatment. ML standard errors were bootstrapped (200 iterations). For two-way fixed effects, standard errors are clustered by household.

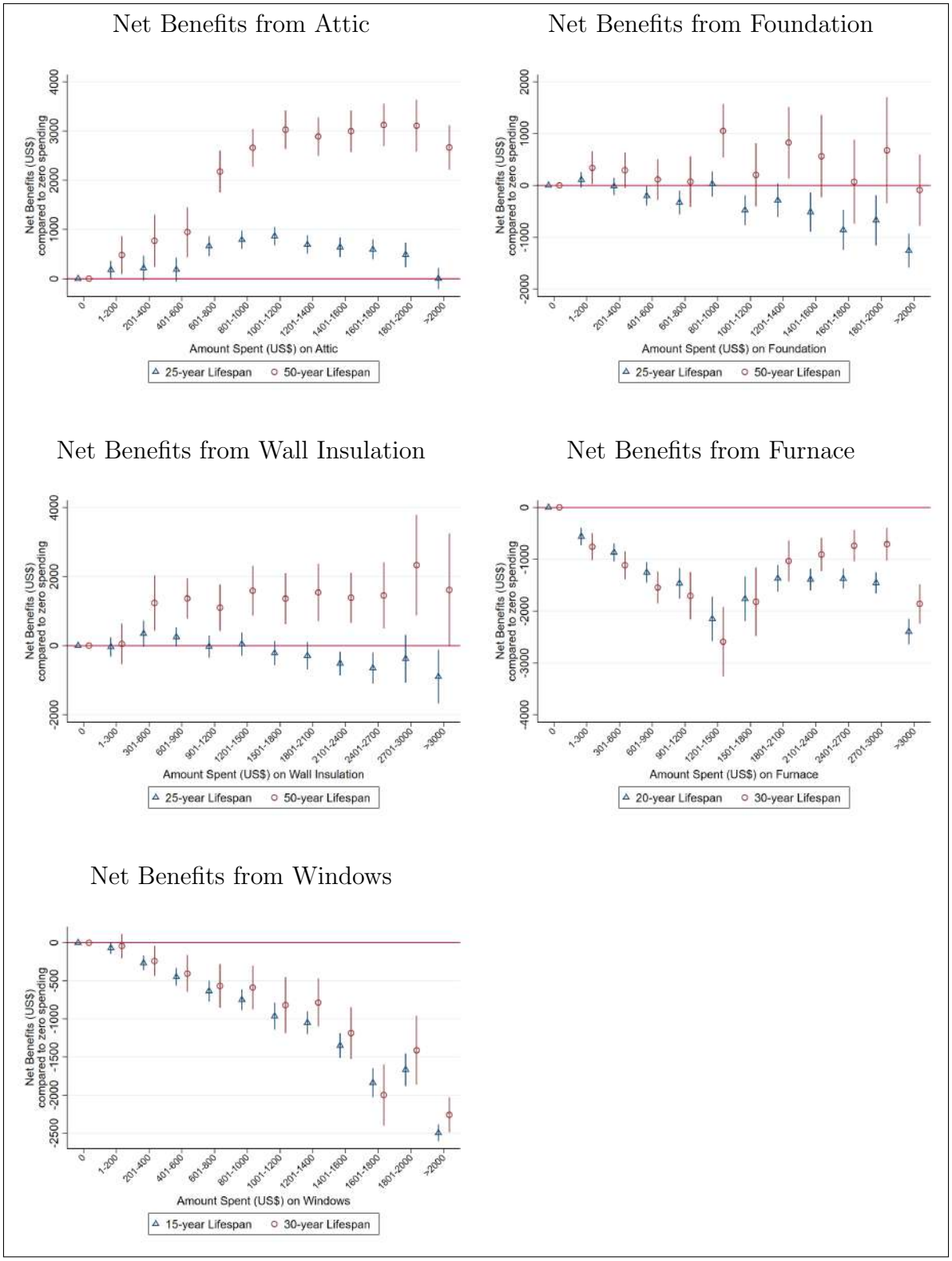


Figure 5: Cost-Effectiveness of Main WAP Spending Categories

Notes: The figures above present results from cost-benefit analyses for the main categories of WAP spending. Blue triangles represent net benefits with baseline assumptions: shorter lifespans, and 3% discount rate. The red circles represent net benefits assuming longer lifespans. Standard errors were bootstrapped (200 iterations).

Online Appendix

A Proofs

Proof of Theorem 1: Based on the Neyman-Rubin potential outcomes framework and Assumption 2, the observed outcome can be written as:

$$Y_{i,t} = Y_{i,t}(0) + D_{i,t}[Y_{i,t}(1) - Y_{i,t}(0)] \iff$$
$$Y_{i,t} - Y_{i,t}(0) = D_{i,t} \cdot b_{i,t}$$

Taking the conditional expectation in both sides, from Assumption 4, we can rewrite the left-hand side in terms of the function $g()$:

$$\mathbb{E}[Y_{i,t} - Y_{i,t}(0)|\mathbf{X}_{i,t}, D_{i,t} = 1] = \mathbb{E}[D_{i,t} \cdot b_{i,t}|\mathbf{X}_{i,t}, D_{i,t} = 1]$$
$$\iff$$
$$\mathbb{E}[Y_{i,t}|\mathbf{X}_{i,t}, D_{i,t} = 1] - g(\mathbf{X}_{i,t}(1)) = \mathbb{E}[b_{i,t}|\mathbf{X}_{i,t}, D_{i,t} = 1]$$

Taking now the conditional expectation for $D_{i,t} = 1$:

$$\mathbb{E}[\mathbb{E}[Y_{i,t}|\mathbf{X}_{i,t}, D_{i,t} = 1] - g(\mathbf{X}_{i,t}(1))|D_{i,t} = 1] = \mathbb{E}[\mathbb{E}[b_{i,t}|\mathbf{X}_{i,t}, D_{i,t} = 1]|D_{i,t} = 1] \quad .$$

Therefore, it follows from Assumption 3 and the LIE:

$$\mathbb{E}[Y_{i,t} - g(\mathbf{X}_{i,t})|D_{i,t} = 1] = \mathbb{E}[b_{i,t}|D_{i,t} = 1] \quad .$$

The proof for the $ATT(r)$ follows by taking all the conditional expectations also with respect to $t - (q_i - 1) = r$, for $r > 0$. Since Assumption 4 holds for all t, it also holds for $t = (q_i - 1) + r$,$r > 0$.

Proof of Theorem 2:

Now, we take the conditional expectation on $\mathbf{X}_{i,t} = \mathbf{c}$ in both sides.

$$\mathbb{E}[Y_{i,t} - Y_{i,t}(0)|D_{i,t} = 1, \mathbf{X}_{i,t} = \mathbf{c}] = \mathbb{E}[b_{i,t}|D_{i,t} = 1, \mathbf{X}_{i,t} = \mathbf{c}]$$

Assumption 4' implies the following:

$$\mathbb{E}[Y_{i,t} - g(\mathbf{X}_{i,t}(1))|\mathbf{X}_{i,t} = \mathbf{c}, D_{i,t} = 1] = \mathbb{E}[b_{i,t}|\mathbf{X}_{i,t} = \mathbf{c}, D_{i,t} = 1]$$

Therefore, it follows from Assumption 3:

$$\mathbb{E}[Y_{i,t} - g(\mathbf{X}_{i,t})|\mathbf{X}_{i,t} = \mathbf{c}, D_{i,t} = 1] = \mathbb{E}[b_{i,t}|\mathbf{X}_{i,t} = \mathbf{c}, D_{i,t} = 1] \quad .$$

Consistency Properties:

The Machine Learning algorithm (XGboost, from Chen and Guestrin, 2016) used in this paper relies on numerical optimization in function space. The optimization minimizes the expected value of a loss function based on the Euclidean distance between the observed outcome and the predicted value from a linear combination of many regression trees. Consistency properties of this algorithm are shown in Biau and Cadre (2021).

Under the Law of Large Numbers and the consistency of $\hat{g}()$, we can show that $\widehat{ATT}$ is a consistent estimator of the ATT.

$$\begin{aligned}
\widehat{ATT} &= \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}\hat{b}_{i,t}\mathbb{1}\{D_{i,t}=1\}}{\sum_{i=1}^{I}(T-(q_i-1))\mathbb{1}\{q_i \leq T\}} \\
&= \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}[Y_{i,t}-\hat{Y}_{i,t}(0)]\mathbb{1}\{D_{i,t}=1\}}{\sum_{i=1}^{I}(T-(q_i-1))\mathbb{1}\{q_i \leq T\}} \\
&= \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}Y_{i,t}\mathbb{1}\{D_{i,t}=1\}}{\sum_{i=1}^{I}(T-(q_i-1))\mathbb{1}\{q_i \leq T\}} - \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}\hat{g}(\mathbf{X}_{i,t})\mathbb{1}\{D_{i,t}=1\}}{\sum_{i=1}^{I}(T-(q_i-1))\mathbb{1}\{q_i \leq T\}} \\
&\xrightarrow{p} \mathbb{E}[Y_{i,t}|D_{i,t}=1] - \mathbb{E}[g(\mathbf{X}_{i,t})|D_{i,t}=1]
\end{aligned}$$

where the term of convergence identifies the ATT according to Theorem 1.

B Proposed Bootstrap Algorithm

As described in section 3.4, for inference with the machine learning estimates, I also need to take into account that there is uncertainty in the predictive step of the method. It is reasonable to assume that the predictive model behaves differently depending on the sample with which it is trained. Chernozhukov, Fernández-Val, and Luo (2018) propose bootstrapping for improving confidence bands in settings with heterogeneity. Therefore, as an alternative for the procedure from equation (6), I propose implementing a conservative algorithm that uses bootstrapped standard deviations of the parameters of interest as an approximation for their standard errors. The bootstrap algorithm can be summarized as follows.

Bootstrap Algorithm: *Let N be the total number of observations in the sample. Let $b = 1...B$ denote a bootstrap iteration. (1) Draw (ω_1, ..., ω_N), which is a vector of N nonnegative bootstrap weights attributed to each observation in the sample. Once those weights are applied to the original sample, a new bootstrapped sample is constructed. To obtain the weights, employ stratified (by home) random sampling with replacement, such that $N_b \approx N$ (i.e., bootstrap sample should be approximately the same size as the original sample). (2) Run the machine learning predictive model with the bootstrap sample and obtain predictions. (3) Transform the predictions (e.g., calculate averages, run regressions), and return the parameter of interest β_b (e.g., WAP treatment effect). (4) Repeat steps 1 through 3 for total of B bootstrap iterations. (5) Compute the bootstrapped standard error as $\sigma = \frac{\sum_{b=1}^{B}(\beta_b - \hat{\beta})^2}{B}$ (i.e., the standard deviation of the parameter of interest across bootstrap samples). (6) Compute confidence bands around the parameter of interest as $\beta_- = \beta - 1.96 \times \sigma$ (lower bound), and $\beta_+ = \beta + 1.96 \times \sigma$ (upper bound).*

Within the simulated setting (Spanish air pollution data), I test the stability of standard errors generated with the algorithm described above. Figure B.1 plots the evolution of standard deviations (SDs) of the ML estimates of ATT in Table 1 Panel A. Substantial instability can be noted during the first 50 iterations. However, the variation in SDs becomes negligible after 60 iterations (ranging between 0.23 and 0.24). Further,

the trend seems to be decreasing, such that machine learning estimates could potentially be even more precise with more iterations. In this context, however, there is a tradeoff between compute time and precision of the estimates.

I similarly test the stability of the bootstrapped standard errors within the real data application (evaluation of the Weatherization Assistance Program). Figure B.2 shows the standard deviations after each bootstrap iteration of ATT estimates for the Program. After 90 iterations, the standard deviations stabilize, ranging from 0.0027 to 0.0028. A comparison of Figures B.1 and B.2 reveals that the optimal number of bootstrap iterations, to reach stability of estimated standard errors, depends on the underlying data structure and on the research context.

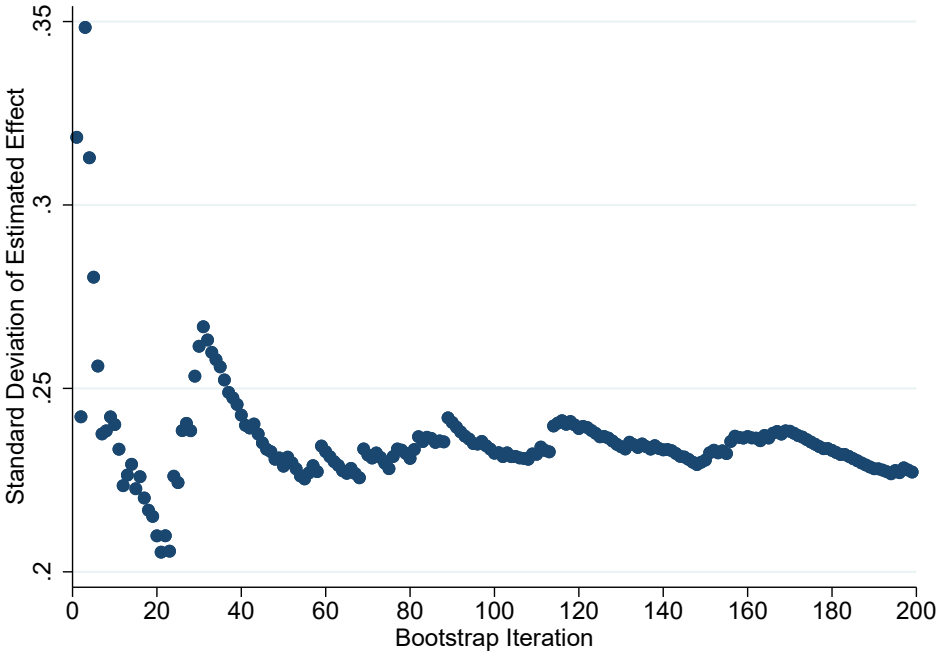


Figure B.1: Stability of Bootstrapped Standard Deviations
(for ML estimates of ATT simulations in Table 1A)

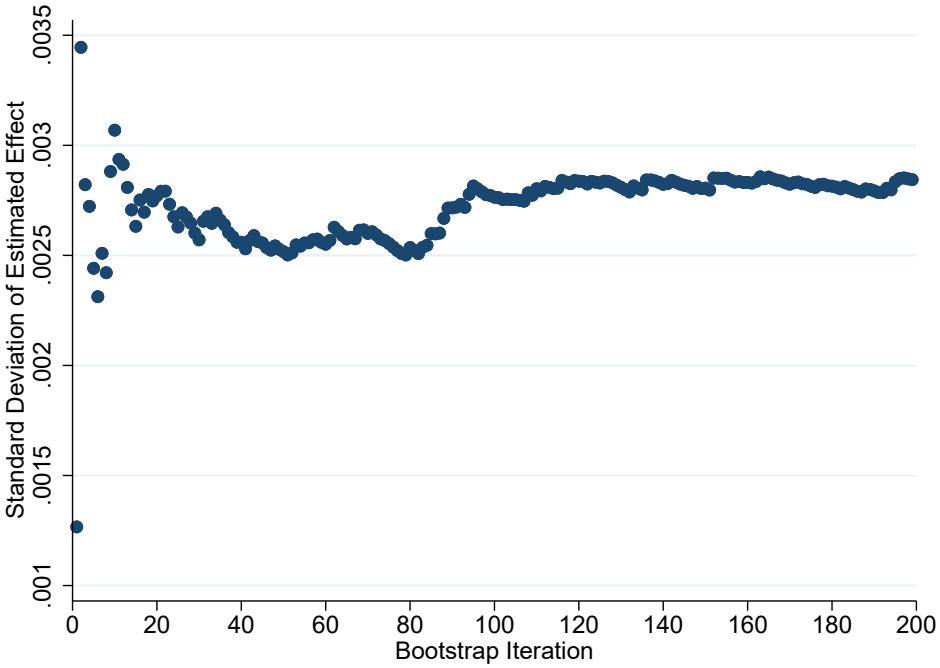


Figure B.2: Stability of Bootstrapped Standard Deviations
(for ML estimates of ATT from the Illinois Weatherization Assistance Program)

C Descriptive Statistics for Data Used in Simulations

Here I present descriptive statistics for the variables used in the simulations from section 4. All data from the simulations are publicly available and will be shared along with the replication packet.

Table C.1: Air Quality and Weather in Spain

	Average	Std Dev	Min	Max
PM_{10} Concentration ($\mu g/m^3$)	21.54	13.88	0.00	691.00
Min Temperature (C)	12.21	6.36	-15.60	28.90
Median Temperature (C)	16.79	6.39	-8.40	35.70
Max Temperature (C)	21.38	7.05	-6.90	45.40
Precipitation(mm)	1.69	5.81	0.00	144.90
Wind Direction (degrees)	19.58	10.55	0.00	36.00
Wind Speed (m/s)	3.25	2.04	0.00	21.90
Max Atmospheric Pressure (hPa)	990.63	28.55	774.80	1044.00
Min Atmospheric Pressure (hPa)	986.13	28.72	769.30	1040.40
Number of Observations	170,484			

Notes: Air pollution data collected from 311 monitors across Spain (MITECO, 2020), with coverage from 2014 to 2019. Several monitors have missing data for some dates, such that this constitutes an unbalanced panel. These were matched with daily weather data from 271 stations across the country, obtained from AEMET (2020).

Table C.2: Air Quality Station Details and Wildfires in Spain

	Average	Std Dev	Min	Max
Station Zone (percent):				
Industrial	5.47	22.77	0.00	100.00
Residential	22.83	42.04	0.00	100.00
Residential/Commercial	25.40	43.60	0.00	100.00
Residential/Industrial	21.22	40.95	0.00	100.00
Other	25.08	43.42	0.00	100.00
Station Location (percent):				
Urban	56.27	49.69	0.00	100.00
Suburban	43.73	49.69	0.00	100.00
Station Type (percent):				
Industrial	35.05	47.79	0.00	100.00
Traffic	27.33	44.64	0.00	100.00
Background	37.62	48.52	0.00	100.00
Number of Stations	311			
Wildfires (hectares burned)	85,512.82	54,195.02	23,911.89	178,482.38
Number of Years	6			

Notes: Details from air quality stations obtained from MITECO (2020). Annual national-level wildfire data from Spain obtained from MITECO (2021).

Table C.3: Electricity Generation in Spain

	Average (MWh)	Std Dev	Min	Max
Total Generation	669,251.58	71,634.19	480,778.91	889,030.50
Hydro	15,484.06	6,225.23	3,883.10	30,592.30
Nuclear	150,882.93	19,711.13	83,788.80	175,915.50
Natural Gas	84,461.27	52,748.13	15,061.10	315,329.81
Wind	135,819.19	72,028.49	17,787.00	406,145.41
Solar PV	21,844.07	7,009.55	4,684.70	34,844.20
Solar Thermal	14,438.02	10,265.74	6.00	33,380.90
Natural Gas Cogeneration	67,403.91	7,486.51	34,026.90	88,428.40
Biomass	9,546.50	2,686.70	4,343.40	15,528.30
Number of Observations	2,191			

Notes: This table presents statistics for national level electricity generation by fuel in Spain. All values are in MWh and are recorded daily. Data obtained from ESIOS (2020).

Table C.4: Demographics in Spain

	Average	Std Dev	Min	Max
Total Population (thousands)	1,087.61	1,178.24	134.14	6,578.08
Male Population	534.74	571.43	67.93	3,147.87
Female Population	552.87	606.91	66.21	3,430.21
Total Employment	438.99	555.04	55.20	3,427.40
Agriculture Employment	19.10	14.51	2.70	72.90
Industry Employment	52.00	67.59	5.60	383.20
Manufacturing Employment	46.18	61.26	4.50	350.00
Construction Employment	25.77	29.33	4.30	187.70
Commerce Employment	141.86	184.59	14.50	1,144.70
Finance Employment	65.18	115.45	4.50	774.70
Public Sector Employment	135.08	167.81	17.20	1,101.30
Total GDP (billion Euros)	25.92	35.72	3.22	230.81
Agriculture GDP	0.76	0.51	0.09	2.50
Industry GDP	3.97	5.52	0.49	31.23
Manufacturing GDP	3.10	4.64	0.26	27.24
Construction GDP	1.39	1.58	0.24	9.95
Commerce GDP	6.31	9.94	0.64	68.98
Finance GDP	5.66	9.28	0.48	64.75
Public Sector GDP	5.41	6.63	0.77	43.55
Number of Observations	175			

Notes: Demographic data recorded annually and at the province level. Data obtained from INE (2020).

Table C.5: Firm Entries and Exits Across Provinces in Spain

	Average	Std Dev	Min	Max
Firm Exits	37.44	60.93	0.00	947.00
Firm Entries	190.63	300.14	3.00	2,118.00
Firms' Capital (million Euros)	10,079.31	24,603.35	24.00	413,001.00
Number of Observations	2,042			

Notes: Firm entries, exits, and total capital at the province-by-month level. Data obtained from INE (2020).

D Machine Learning Model Tuning and Diagnostics – Simulations

To predict air pollution (PM_{10}) concentrations, I supply the above control variables to a machine learning algorithm called XGBoost, which is a computationally fast implementation of gradient boosted trees, developed by (Chen and Guestrin, 2016). Consistency properties of this algorithm are shown in Biau and Cadre (2021). The concept of boosted trees involves iteratively combining 'weak' predictive trees to form an ensemble. Each tree is constructed with a fraction of the set of the available control variables. More weights are given to the trees with better predictive accuracy. By default, the algorithm uses mean squared errors (MSE) as a measure of accuracy. With this algorithm, a researcher can therefore be agnostic in terms of which variables to include for prediction, as well as their functional forms. Note that regression trees intrinsically consider variable interactions and binning. As the tree "depth" increases, interactions become more complex. With more tree "branches," I allow for more flexibility in how each variable is included.

To increase predictive accuracy of machine learning models, it is common practice to "tune" the (hyper)parameters that control factors such as maximum tree depth. The following section describes the configurations that I considered for the model.

D.1 Hyperparameter Tuning – Simulations

Prior to settling on a model that performs well in terms of predictions, I perform hyperparameter tuning via 5-fold cross-validation. This was implemented through the "SuperLearner" package in R (Polley et al., 2018). Sample splits for the validation folds

are random.[30] I consider the variations to following XGBoost hyperparameters:

- Number of trees/iterations: determines the total number of models of which the ensemble XGBoost is constituted. (either 2000 or 3000)
- Maximum tree depth: correlated with the complexity of the model and variable interactions. (either 10 or 30)
- Shrinkage/step-size/learning-rate/eta: a rate between 0 and 1, that determines the contribution of each new tree to the ensemble. Lower values are more conservative and prevent overfitting. (set at 0.05)
- Minimum observations per node: correlated with the frequency of branch splits, which also determines the sizes of bins considered for each variable. Smaller nodes imply more flexible models, but may also lead to overfitting. (either 20 or 60)

Other XGBoost hyperparameters were set at their defaults. Therefore, I test a total of 8 hyperparameter configurations. Table D.1 below presents performance diagnostics for each hyperparameter combination. Model ID 2, highlighted in gray, was the best-performing one, with a RMSE of 6.524.

[30] In panel data settings one may consider "vertical" or "horizontal" cross-validation. For the simulations in this paper, vertical CV implies stratifying sample splits by air quality station, while horizontal CV means splitting across time. For ex-post evaluation settings, I recommend the latter. Stratification can lead to overfitting the model for stations that are in the training set, such that accuracy will be lower in the validation set (constituted of completely "unseen" stations). However, for this paper, "out-of-sample" is defined as an unseen set of dates (as opposed to stations), such less biased prediction errors can be obtained by splitting across time. Athey, Bayati, et al. (2019) provide a more complete discussion of vertical versus horizontal cross-validation. See Arlot and Celisse (2010) for a survey on cross-validation techniques.

Table D.1: Hyperparameter Tuning – Simulated Setting

Model ID	Number of Trees	Max Tree Depth	Min Obs per Node	Shrinkage	In Sample RMSE	Cross-Validated RMSE
1	2000	10	20	0.05	2.793	6.574
2	3000	10	20	0.05	2.149	6.524
3	2000	30	20	0.05	0.203	6.852
4	3000	30	20	0.05	0.068	6.853
5	2000	10	60	0.05	3.818	6.686
6	3000	10	60	0.05	3.248	6.601
7	2000	30	60	0.05	1.029	6.618
8	3000	30	60	0.05	0.598	6.627

Notes: Performance metrics for XGBoost (Chen and Guestrin, 2016) algorithms for prediction of PM_{10} air pollution concentrations measured in $\mu g/m^3$. Control variables are presented in Appendix C. As a reference, the RMSE may be compared to the standard deviation of the outcome variable, in this case equal to 13.88 $\mu g/m^3$. The best-performing model (Model ID 2) is highlighted in gray.

D.2 Prediction Errors by Covariates – Simulations

Figure D.1 presents in-sample versus cross-validated prediction errors (residuals) by selected covariates used in the simulations. These were obtained by running equation (4) from the main text. As already shown in Figure 1, in-sample residuals are smaller, thus potentially masking some sources of biases. Nevertheless, cross-validated errors in this setting are close to zero throughout almost all tested bins. Importantly, errors by altitude bins (mid-right panel) are not significantly different from zero, which supports the heterogeneity analyses from the simulations in section 4.2.

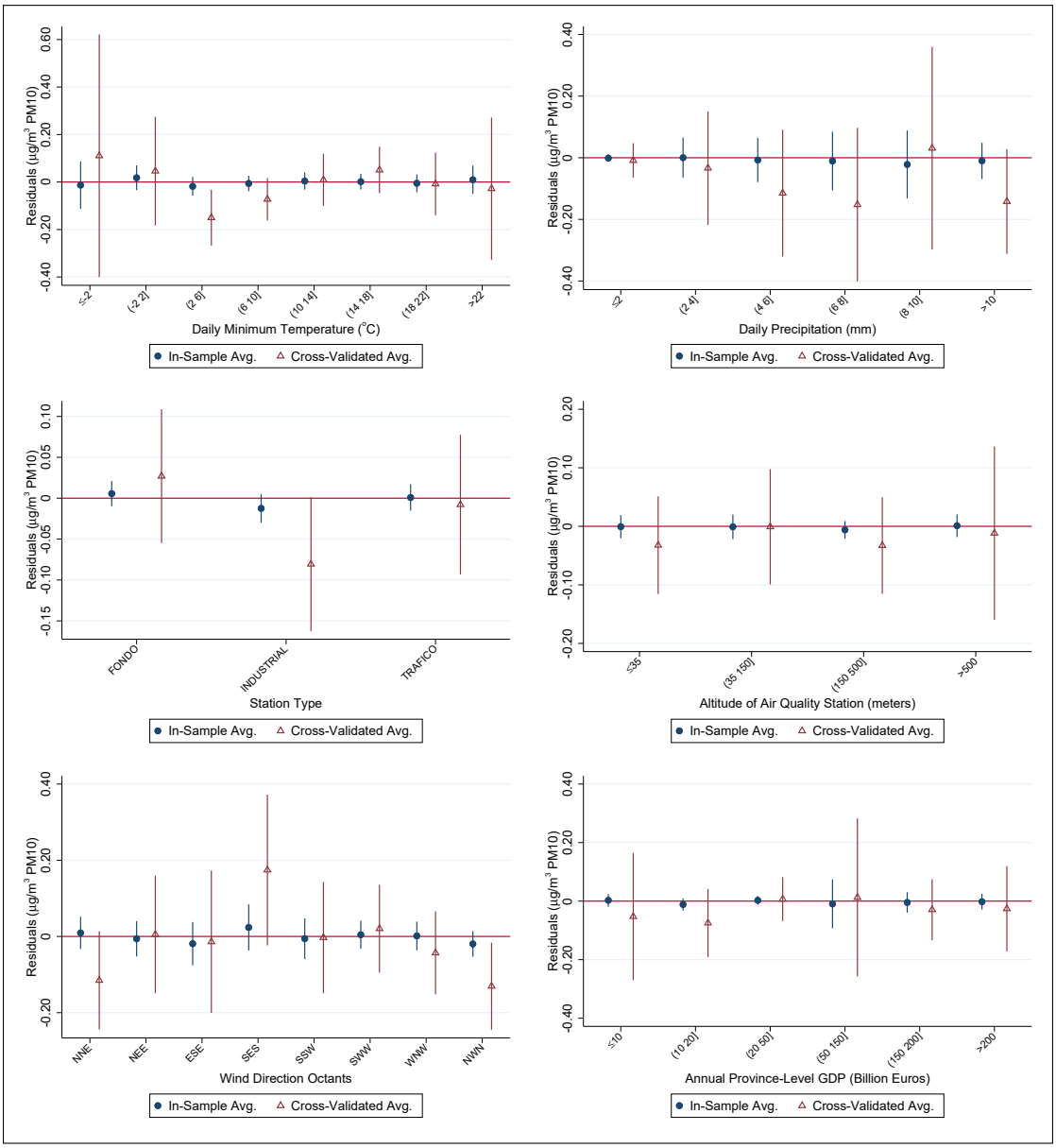


Figure D.1: Simulation Prediction Errors by Selected Covariates

E Summary Statistics for the Weatherization Assistance Program Sample

Table E.1 presents descriptive statistics for the main variables collected during the Weatherization Assistance Program application process and pre-treatment home energy audits. In terms of demographics, it can be noted that the sample of treated households is constituted primarily of low-income families (average yearly income around $17,220). They are also mostly middle aged (∼54 years) homeowners (94%). The variables related to housing structure reveal that very diverse homes are weatherized by the program: there is significant variation in floor area, pre-treatment blower door tests, number of bedrooms, and even vintage.

Figure E.1 represents the histogram of pre-teatment natural gas usage for homes served by the program. The average usage is around 11 MMBtu, but with significant variation. Notably, a lot of the distribution is concentrated at lower levels, likely during summer of warmer months when natural gas is not needed so much.

Table E.1: WAP Descriptive Statistics

	Average	Standard Deviation	Min	Max
Income($/1000)	17.32	10.33	0.00	52.48
N Occupants	2.97	1.73	1.00	9.00
Householder Age	54.83	15.54	22.00	89.00
Female Householder (%)	0.66	0.47	0.00	1.00
Renter (%)	0.06	0.23	0.00	1.00
Seniors 65+ (%)	0.39	0.49	0.00	1.00
Children Under 18 (%)	0.17	0.38	0.00	1.00
Blower Door Pre (CFM50)	3645.46	1662.58	980.00	13662.00
Heating Unit Size (kBTU)	87.10	38.56	0.00	150.00
Floor Area (sqft)	1543.70	600.28	600.00	3774.00
N Bedrooms	4.74	0.74	1.00	5.00
N Windows	16.91	5.73	2.00	26.00
Has Multiple Stories (%)	0.45	0.50	0.00	1.00
Number of Homes	34,497			

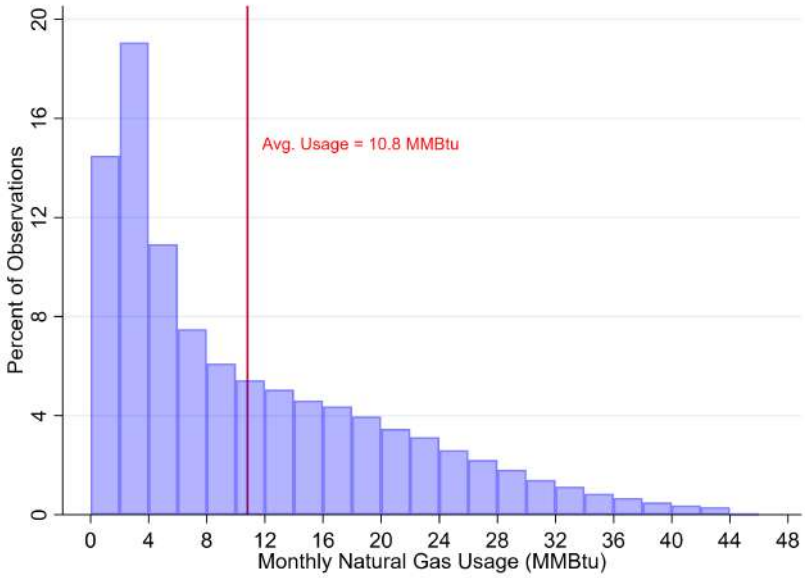


Figure E.1: Histogram of Pre-Treatment Energy Usage

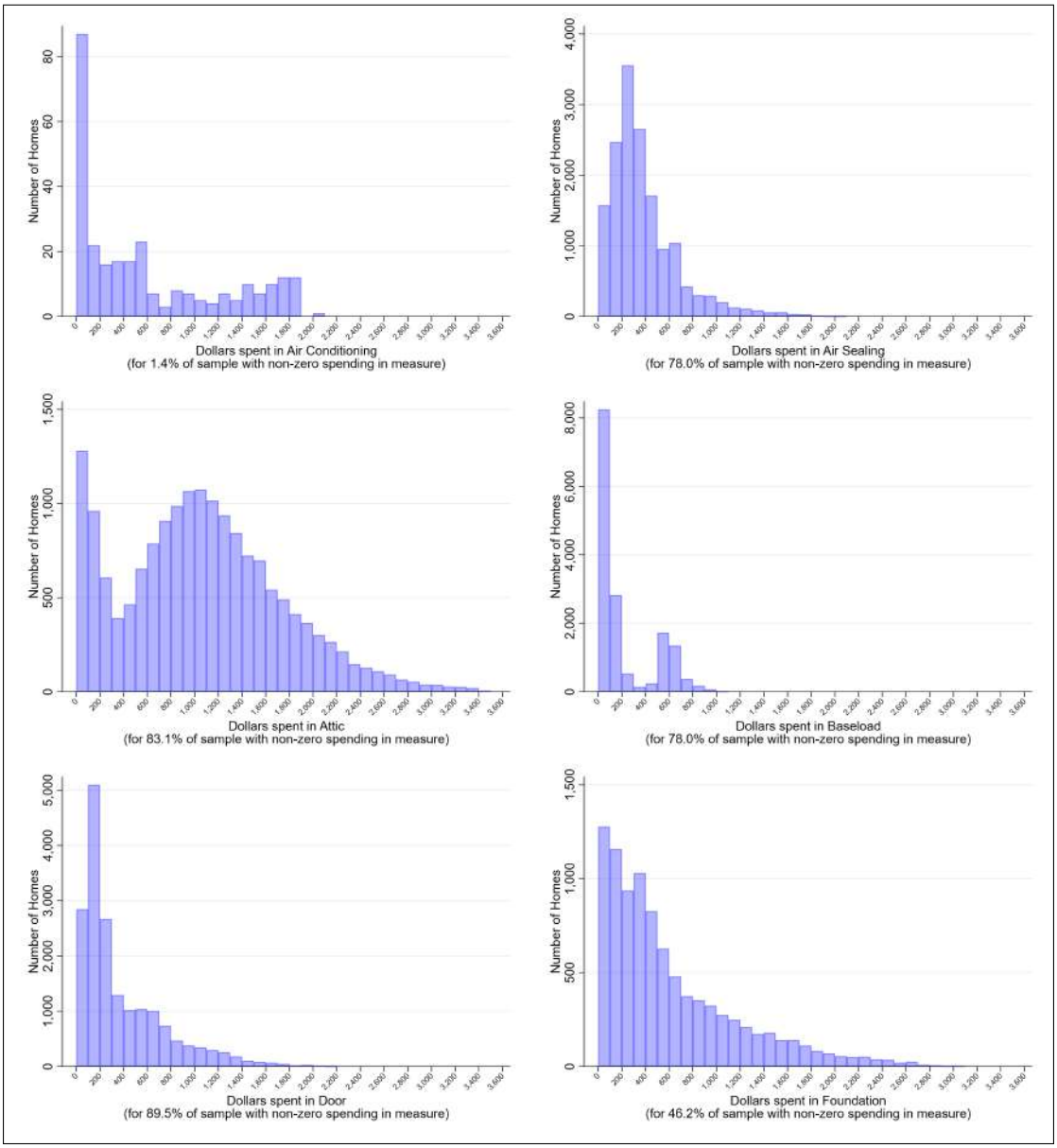


Figure E.2: Histograms for Categories of WAP Spending

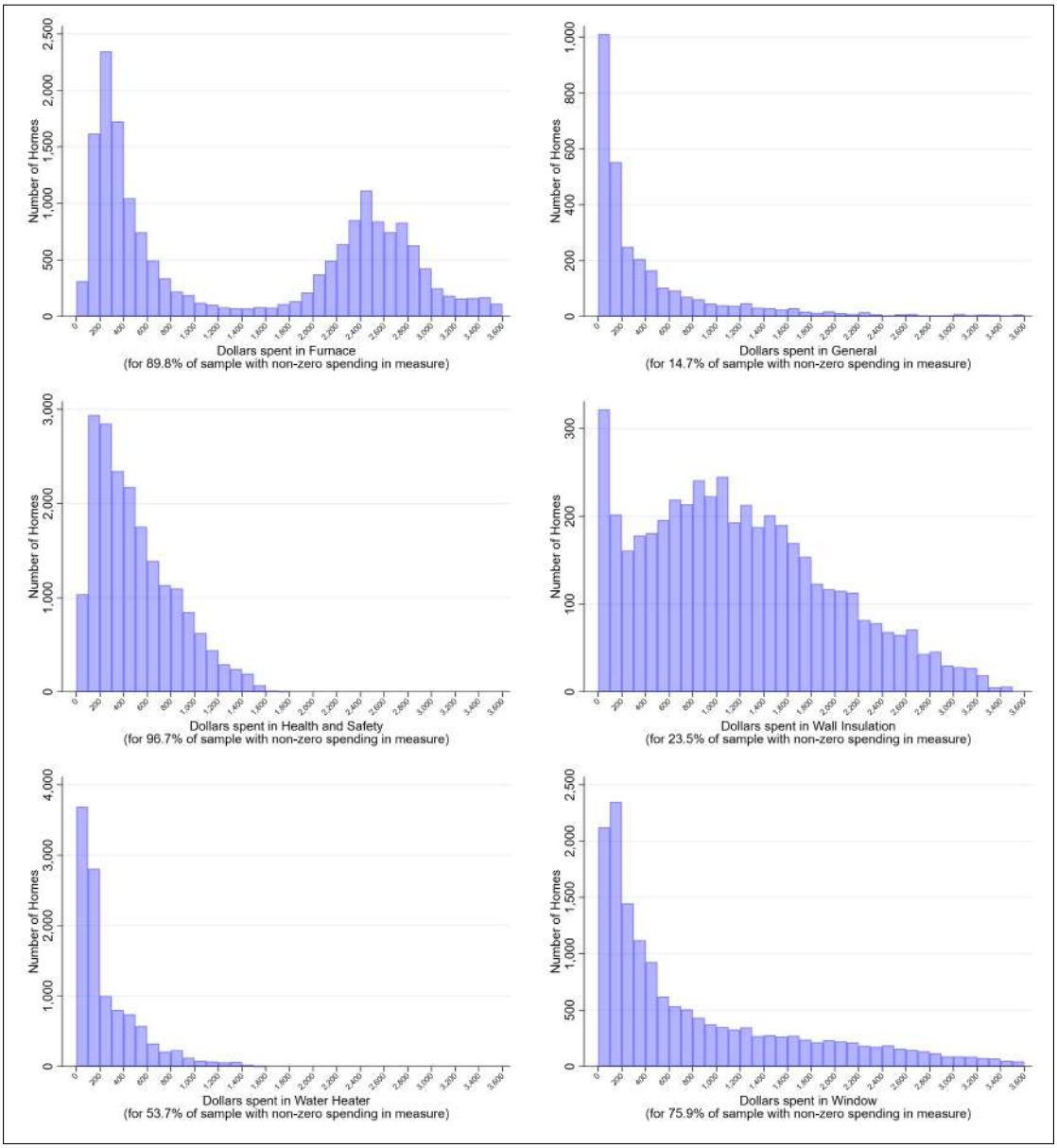


Figure E.2 (cont.): Histograms for Categories of WAP Spending

F Machine Learning Model Tuning and Diagnostics – Real Data Application

The machine learning model for the real data application (WAP sample) was trained only with data available prior to weatherization, namely: pre-treatment billing data, energy audit information, housing structure variables, household demographics, and weather variation. Specifically, I include the following variables: energy usage in MMBtu (outcome), min. outdoor temperature, max. outdoor temperature, precipitation, floor area (square feet), family size, number of windows, number of stories, number of bedrooms, vintage, county indicator, building shielding class (measure of shielding provided by structures surrounding home), pre-treatment blower door test (CFM50), main heating system type, main heating system capacity (Btu), attic R-value, household income, indicators for householder's race, presence of disable occupant, presence of children, presence of elderly, home priority rank, audit date (month, year, and day), program year of audit, month of year, year of sample, number of days in billing cycle, monthly average natural gas prices in Illinois, and monthly average electricity prices in Illinois. The outcome (natural gas usage) varies by home and by month of sample (billing period). Weather also varies by month of sample, while information collected during WAP audit/application varies only across homes.

F.1 Hyperparameter Tuning – Real Data Application

For the WAP real data application, I also focus on gradient boosted trees (XGBoost; Chen and Guestrin, 2016). Diagnostics presented in Table F.1 below were obtained via 5-fold cross-validation, with sample splits defined at random (not stratified).

Table F.1: Results from Hyperparameter Tuning – Real Data Application

Model ID	Num. Trees (Iterations)	Max. Tree Depth	Shrinkage	Min. Observations per Node	Mean Squared Error	Ensemble Weight
1	1000	20	0.05	30	14.144	0.475
2	2000	20	0.05	30	14.232	0.000
3	1000	30	0.05	30	14.148	0.466
4	2000	30	0.05	30	14.227	0.000
5	1000	20	0.5	30	17.477	0.000
6	2000	20	0.5	30	17.477	0.057
7	1000	30	0.5	30	17.686	0.000
8	2000	30	0.5	30	17.686	0.003

The second-to-last column of Table F.1 reports the mean-squared errors according to each configuration. Rather than choosing a single configuration, for this application the selected machine learning algorithms is an 'ensemble' model which combines predictions across several configurations. The SuperLearner R package (Polley et al., 2018) automatically builds the ensemble, giving higher weights (based on non-negative least squares) to the configurations with lowest MSE. Results suggest that models with lower learning rate (shrinkage = 0.05) were generally more accurate. Further, the ensemble seems to favor less complexity (number of trees = 1000). Note that model IDs 1 and 3 have the highest weights and constitute 94% of the ensemble.

F.2 Prediction Errors – Real Data Application

I test Assumption 2 (no anticipatory effects) and Assumption 4 (stability of the counterfactual function) within the real data setting. For that, I regress cross-validated residuals on indicators for time relative to treatment, as described in equation (3) of the main text. Results are presented in Figure F.1. Note that the coefficients are no larger than 0.04 MMBtu and no smaller than -0.05 MMBtu. This attests to the remarkable predictive performance of the ML algorithm. Further, when evaluated at the 5% or 1% significance level, an F-test rejects the joint significance of the coefficients.

Figures F.2 and F.3 present ML in-sample and cross-validated residuals plotted against bins of monthly energy consumption on the horizontal axis. I note that the model performs extremely well in general, with in-sample residuals generally not greater than 0.5 MMBtu. The cross validated residuals are also small, except for months when gas usage is above 30 MMBtu. But for those cases, errors in percentage point temrs can also be considered small. Further, the graph also shows that those are sparse regions of the sample.

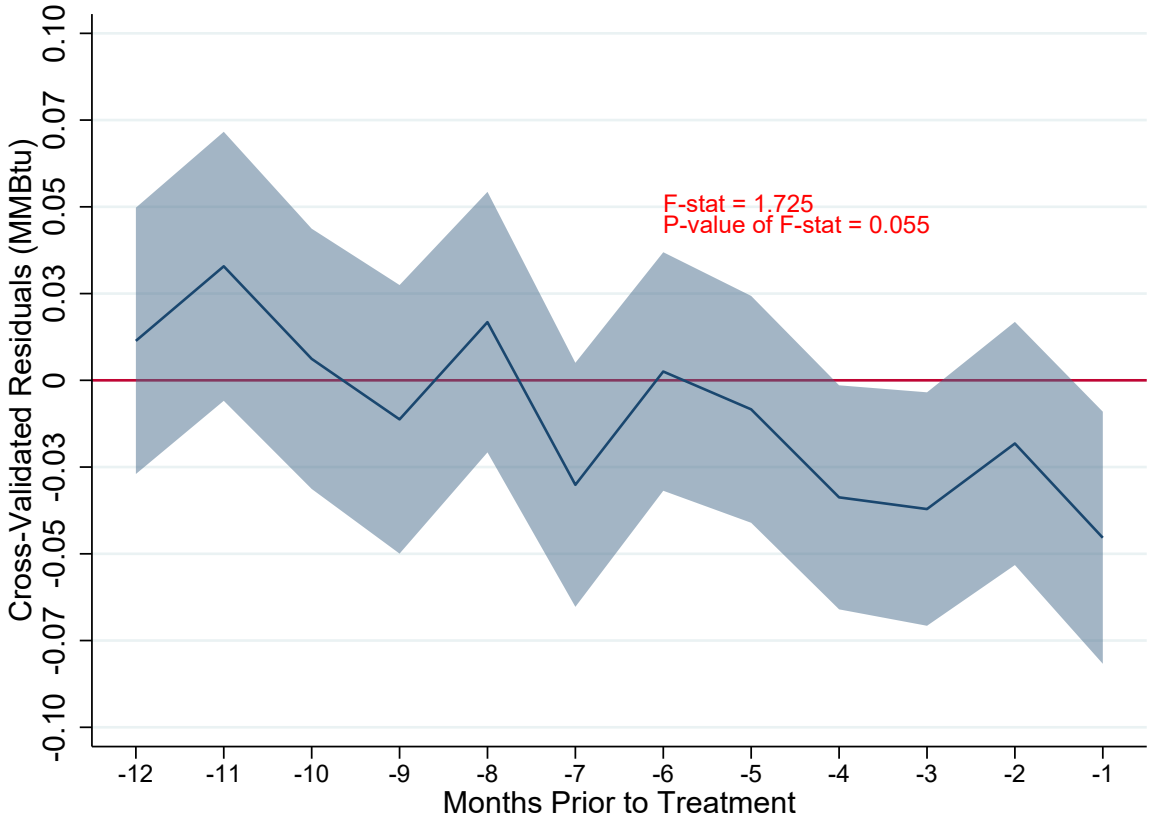


Figure F.1: Assessing Anticipatory Effects and the Stability of the Counterfactual Function – Real Data Application

Notes: This Figure plots coefficient estimates and 95% confidence intervals from a regression of cross-validated residuals on indicators for time relative to treatment (equation 3) for the real data application. This is for testing Assumption 2 (no anticipatory effects), and Assumption 4 (stability of the counterfactual function). This is analogous to "pre-trends" tests in traditional difference-in-differences settings. The top right corner of the Figure shows the resulting F-statistic and associated P-value for a test of joint significance of the coefficients.

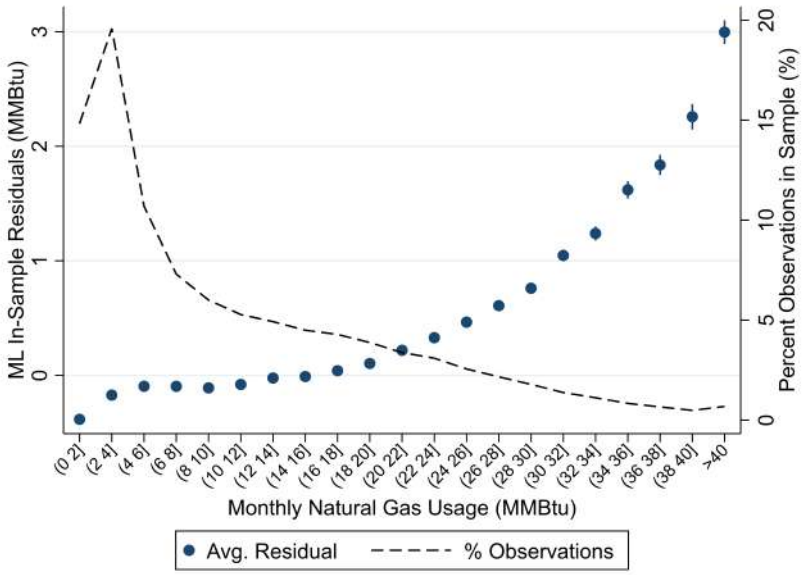


Figure F.2: In-Sample Pre-Treatment Residuals (MMBtu) - real data

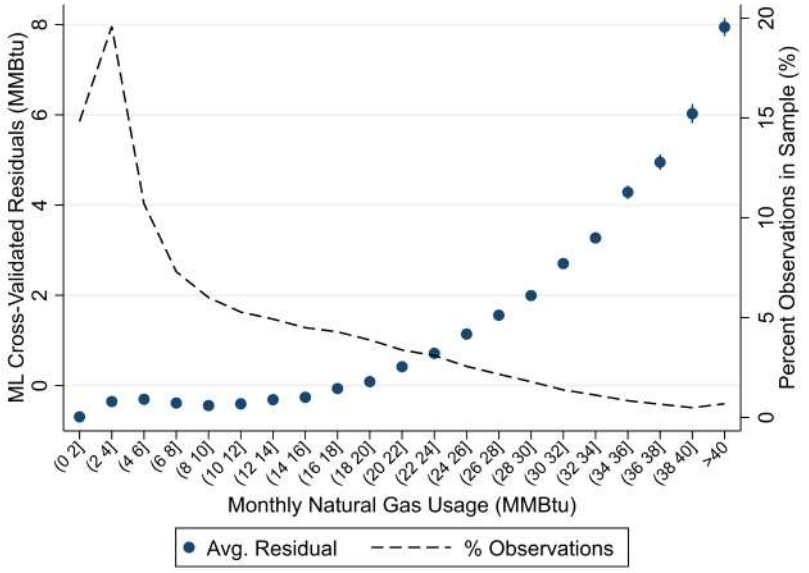


Figure F.3: Cross-Validated Pre-Treatment Residuals (MMBtu) - real data

The following Figure F.4 also shows that prediction errors are not correlated with any of the covariates that are relevant in this context. Again, these were produced with real Weatherization Assistance Program data.

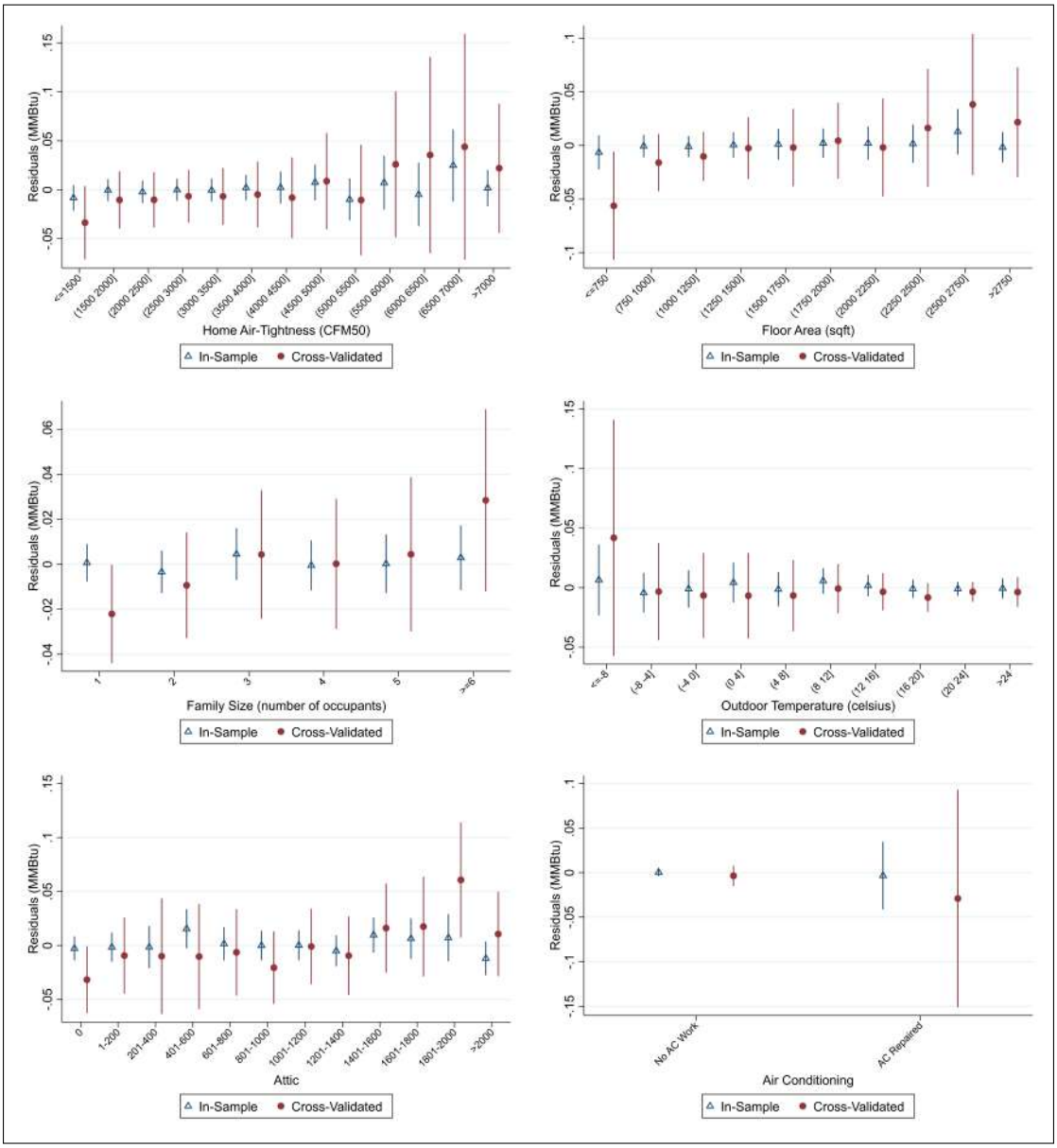


Figure F.4: Cross-Validated Pre-Treatment Residuals (MMBtu) By Covariates

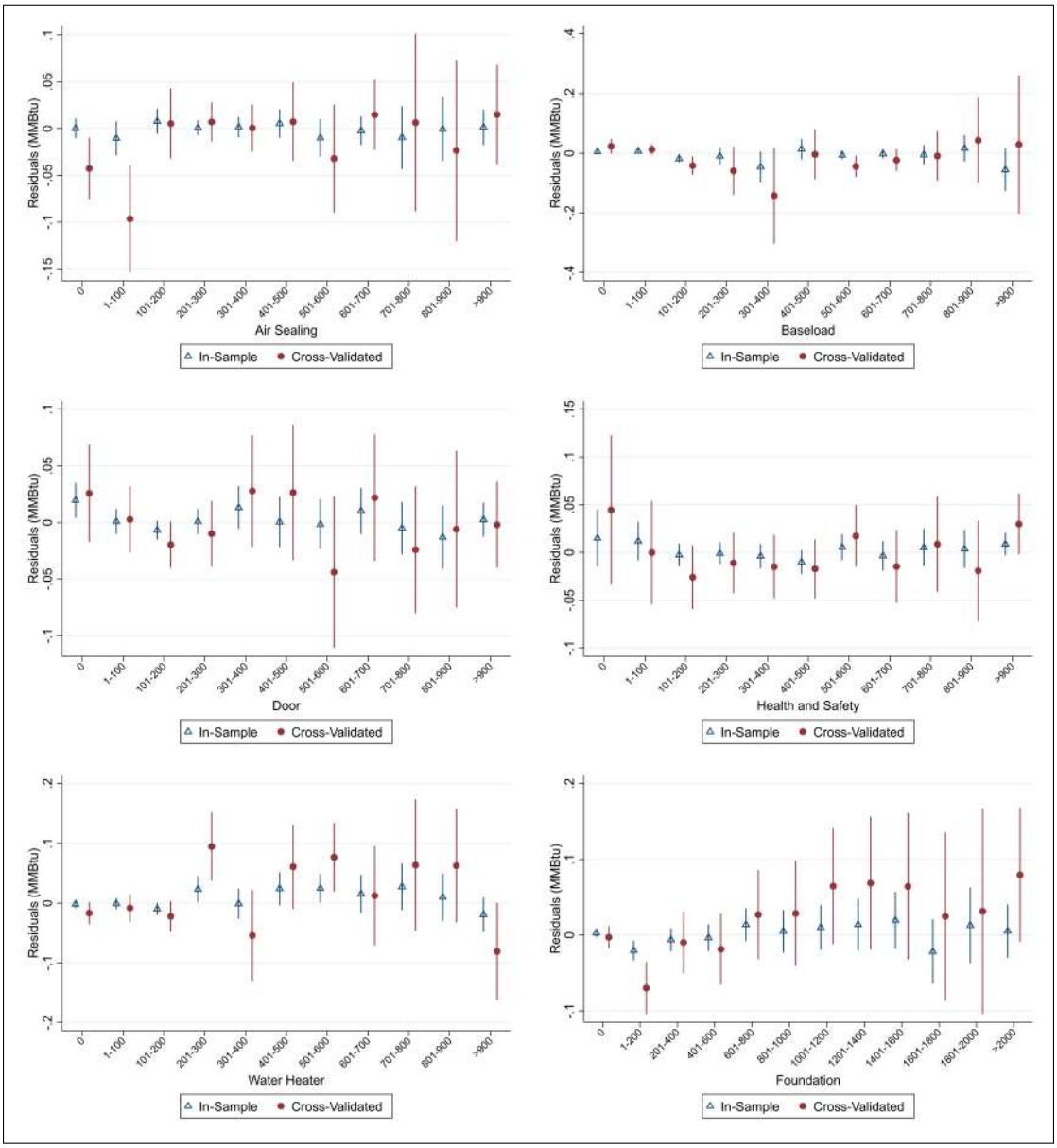


Figure F.4 (continued): Cross-Validated Pre-Treatment Residuals (MMBtu) By Covariates

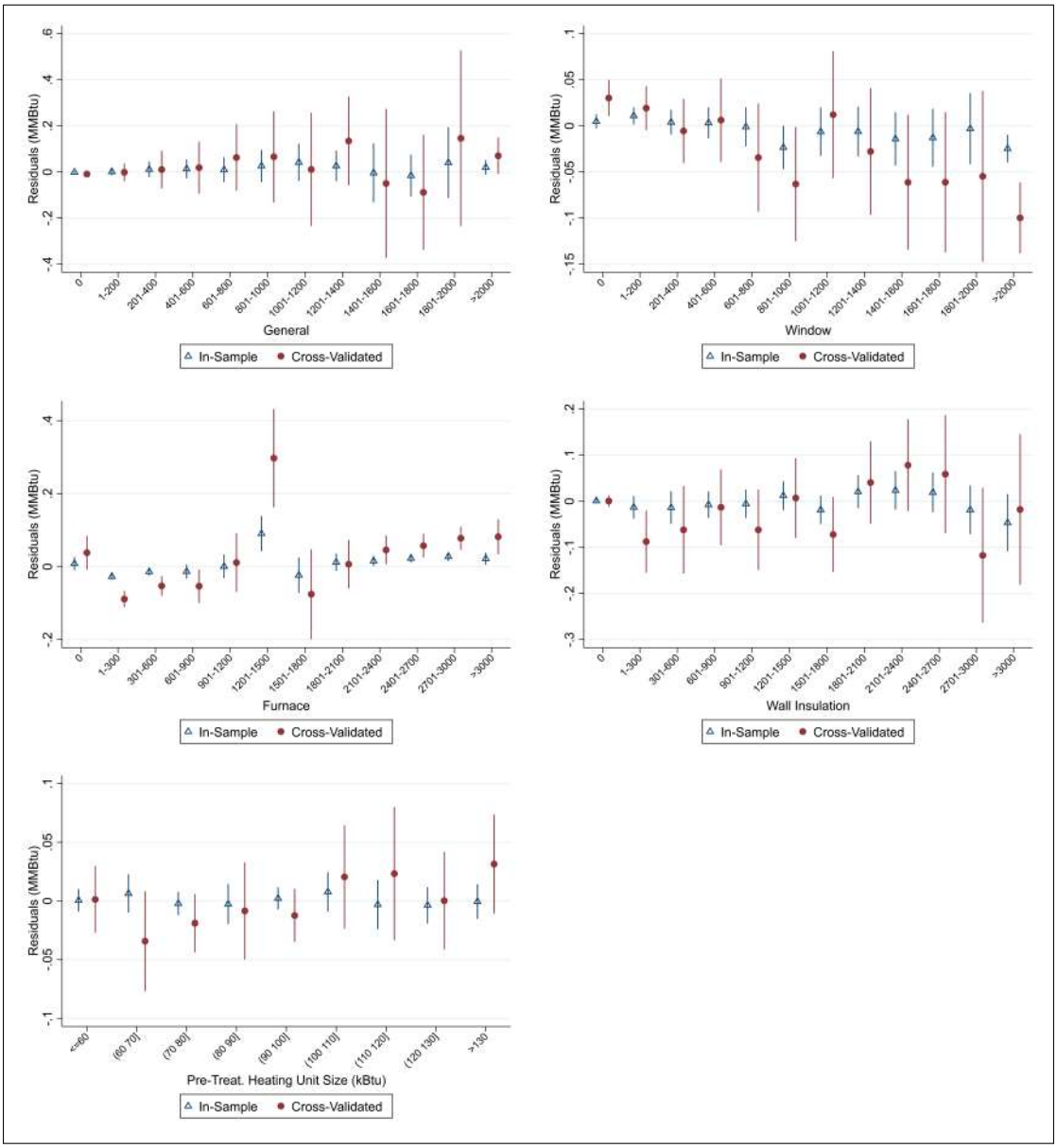


Figure F.4 (continued): Cross-Validated Pre-Treatment Residuals (MMBtu) By Covariates