Testing the validity of machine learning counterfactual prediction methods: An evaluation of ML-generated counterfactuals in low-frequency data conditions

Josefina C Colantoni
Advisor: Dr. Maria Bernedo Del Carpio

Abstract

Researchers who have access to data from only a single treated group currently have limited options for attempting causal inference. Appropriate comparison groups, which can be useful in isolating the effect of a program, are oftentimes unavailable, as in the case of universal policy implementation. A possible alternative to current single-group designs exists in machine learning (ML), which has only recently gained traction as a tool in social science. New developments in ML have resulted in improvements in predictive accuracy that make artificial counterfactual prediction a viable technique in cases of treated-only data; however, evaluations of the internal validity of this technique are lacking: only one exists, and it is in the context of energy use with highly granular, hourly data. This study uses XGBoost, a popular ML algorithm, along with lower-frequency, monthly data from the treated households of an RCT evaluating the effect of water-efficient technologies in rural Costa Rica in order to create household-by-month predictions for what water consumption would have been in the post-treatment period if treatment had not occurred. These artificial counterfactuals are used to derive a treatment effect estimate, which is then compared, in within-study comparison (WSC) fashion, against the treatment effect estimate from the RCT. I find that the ML counterfactual prediction method is able to produce a treatment effect estimate with the same sign as the experimental one, and that is considered equivalent according to a range of popular correspondence measures. However, comparison against alternative single-group designs reveals that a parallel counterfactual prediction approach using OLS as the predictive model rather than XGBoost is able to produce an even closer estimate, suggesting that a simpler model may be more appropriate.

1. Introduction

While randomized control trials (RCTs) are the theoretically preferred research design for causal inference in program evaluation, they are often difficult to implement in practice due to rigid design requirements, ethical concerns, time constraints, and budgetary limits. Consequently, researchers have developed quasi-experimental (QE) methods such as difference-in-difference (DID), comparative interrupted time series (CITS), and regression discontinuity (RDD) to make causal analysis possible in situations where RCTs are not feasible, yet these types of "natural experiments" rely on the fortuitous existence of an appropriate comparison group. Unfortunately, such a group may be unavailable, especially in cases of uniform policy implementation where a treatment is applied to an entire population.

While single-group designs such as the pretest-posttest design and single interrupted time series (SITS) do exist, and require no comparison group, these methods tend to have low internal validity because they are especially vulnerable to bias (Ambroggio et al., 2012). One way internal validity can be assessed is through design replication studies, or within-study comparisons (WSCs), in which a researcher compares the treatment effect estimate derived from an RCT against the treatment effect estimate derived from a QE design that uses the same target population (Wong & Steiner, 2018). WSCs of existing single-group designs have not been able to establish frequent validity (Baicker & Svoronos, 2019).

An emerging, alternative option for researchers who require a single-group design is machine learning (ML). Due to its ability to make accurate predictions, ML has many applications in a broad range of industries including advertising, finance, and healthcare; for this same reason, it is a promising tool for scholars in economics and public policy who are often limited to observational data by circumstance, but still require credible counterfactuals in order to estimate treatment effects (Mohri et al., 2012; Varian, 2014). In situations where high-quality data is available, a researcher can direct an ML algorithm to perform a supervised learning task, during which the algorithm trains a model on a subset of labeled inputs by learning the relationships between covariates and the desired output (Mohri et al., 2012). The researcher can

then use that model to create predictions about the outcomes of another subset of unlabeled inputs. In training a model on pre-treatment data, and using the fitted model to generate predictions about post-treatment outcomes, a program evaluator may generate counterfactuals that can then be used to estimate a treatment effect.

In this paper I use XGBoost, a popular ML algorithm, to generate predictions for post-treatment monthly water use for the treated households of an RCT *had they not received the treatment.* These households in rural Costa Rica were randomly selected to receive water-efficient shower heads and faucet aerators that reduce the volume of water expelled per minute. My ML model is trained on pre-treatment monthly water use, household characteristics, and weather. Using the trained model, I then create a prediction for each household's post-treatment water use, which serves as my counterfactual. I use these artificial counterfactuals in a fixed effects model to produce an estimate of the effect of the installation of the water-efficient technologies on monthly water consumption. Finally, using a WSC framework, I compare this treatment effect estimate against the experimental one in order to assess the validity of this emerging method. If able to replicate experimental results, this ML-generated counterfactual method could present a favorable option to researchers who have access to data from a single treated group only.

Using the Steiner & Wong (2018) correspondence test as my primary measure of similarity, I find that the treatment effect estimate resulting from the ML counterfactual prediction method corresponds to the experimental estimate. This conclusion holds under several alternative methods of correspondence testing; however, comparison against alternative single-group designs reveals that a non-ML counterfactual prediction approach using OLS as the predictive model is able to produce an ATE estimate even closer to the experimental one.

In the following section I present a summary of the existing WSC and ML counterfactual prediction literature. In Section 3 I describe the data used for analysis. In Section 4 I explain the empirical methods used, including a description of the ML counterfactual prediction method, treatment effect estimation, and WSC design. In Section 5 I detail my results, and in Section 6 I discuss their interpretation and implications. Section 7 concludes.

2. Previous Literature

2.1 Within-Study Comparisons

LaLonde is credited with performing the first design replication study. His 1986 evaluation of the National Supported Work Demonstration (NSW) assessed the ability of quasi-experimental techniques using several non-equivalent comparison groups (NECGs) to replicate the results of an RCT. In comparing his quasi-experimental treatment effect estimates against an RCT-derived "benchmark" estimate, LaLonde originated an approach that researchers still use to test the validity of other quasi-experimental study designs.

2.1.a Evaluations of Two-Group Research Designs

Design replication studies, or within study comparisons (WSCs), as they are now more commonly known in the associated literature, have come a long way since they were first introduced by LaLonde. In Wong et al.'s (2018) review of recent WSCs, the authors identified sixty-six such studies performed between 1986 and 2017. These studies evaluated the use of quasi-experimental techniques in a wide range of settings, including development, education, environment, job training, and health (Wong et al., 2018). The majority of the WSCs assess quasi-experimental techniques with comparison groups such as DID, CITS, or RDD (Wong et. al, 2018).

Generally, these two-group techniques have been shown to be able to replicate experimental results[1] when implemented correctly[2] (Cook et al., 2008; St. Clair et al., 2014; St. Clair et al., 2016; Cook et al., 2020; Coopersmith et al., 2022). Two-group designs are appealing from the standpoint of internal validity because they include comparison groups (Shadish et al., 2002); however, there are many situations in which

[1] While each cited author found that one or more of their chosen quasi-experimental method(s) were able to replicate the corresponding RCT benchmark in the context of their specific study, there is currently no field-wide standard for how close the quasi-experimental estimate must be to conclude correspondence. See section 4.3.b for a more detailed discussion of correspondence criteria and measures.

[2] See section 4.3.a for a discussion of the criteria for a successful WSC.

a researcher will not have access to an appropriate comparison group, as with universal policies.

2.1.b Evaluations of Single-Group Research Designs

In instances where a comparison group is not available, researchers currently have limited options. Two such options are the pretest-posttest design and single (or simple) interrupted time series (SITS). These single-group techniques seem to be less popular in practice than two-group designs, and, as a result, there are fewer WSCs involving these methods. The infrequent use of single-group techniques is likely because these methods are inherently susceptible to additional biases that two-group techniques are designed to control for (Shadish et al., 2002; Fretheim et al., 2015; St. Clair et al., 2016). Not surprisingly, of the four WSCs I was able to identify that compared the results of SITS to an RCT benchmark, two found instances where SITS did not produce concordant estimates of the treatment effect.[3]

The major threat to internal validity that single-group designs face is history, or the possibility that events other than treatment could have affected the outcome of interest in the post-treatment period (Shadish et al., 2002; Fretheim et al., 2015). While the internal validity of the ML artificial counterfactual prediction method I will discuss is still susceptible to this threat, the resulting estimates have the potential to improve on SITS because one need not make the assumption that the pre- and post-treatment time trends are modeled by a linear combination of parameters (Baicker & Svoronos, 2019).

2.2 Machine Learning

Varian (2014) espouses machine learning (ML)'s potential as a tool for economists, who are often concerned with uncovering relationships, working with large datasets, and/or making predictions. Predictions generated by novel ML methods are of especial interest because certain algorithms allow for flexible, nonlinear interactions

[3] Both of the studies that found SITS *was* able to replicate RCT results were drug trials (Fretheim et al., 2013; Shadish et al., 2016), and one, Shadish et al. (2016), used data from only 6 cases, which is a much smaller sample size than is typical for econometricians. The only WSC of SITS I could find in social science literature found that SITS did not reproduce the RCT results (Baicker & Svoronos, 2019).

between covariates (Varian, 2014), and these nonlinear models often have greater predictive accuracy than linear models (Amiri et al., 2020; Prest et al., 2023; Chen, 2021).

2.2.a ML Counterfactual Prediction Methods and Model Selection

Recent papers have capitalized on the accuracy improvements offered by ML and have used ML prediction methods as a way to generate counterfactuals. While some of these studies use linear models from algorithms such as LASSO or matrix completion to generate their predictions (Burlig et al., 2020; Dueñas et al., 2021; Athey et al., 2021), many others use algorithms that create nonlinear models such as neural networks, random forests, LightGBM, or XGBoost (Hartford et al., 2016; Christiansen et al., 2021; Souza, 2022; Zhang et al., 2022; Prest et al., 2023). Models produced by the latter algorithms—neural networks, random forests, LightGBM, and XGBoost—are considered non-interpretable models (Weller et al., 2021).

A common qualm economists have about ML is that the resulting models can be somewhat of a "black box" and, unlike traditional regression models, "don't offer simple summaries of relationships in the data" (Varian, 2014). This is true: in choosing ML, one must often sacrifice some degree of interpretability; however, a tradeoff typically exists between interpretability and accuracy. In studies that use ML counterfactual prediction methods, for which the researcher is concerned primarily with predictive accuracy rather than the interpretation of any one coefficient of group thereof, an accurate—although non-interpretable—model is appropriate (Raschka & Mirjalili, 2019; Weller et al., 2021).

2.2.b WSCs of ML-Generated Counterfactuals

Because more accurate counterfactual predictions are able to yield more accurate program effect estimates (Varian, 2014), ML-generated artificial counterfactuals may provide a promising alternative to researchers who wish to attempt causal inference in instances where appropriate comparison group data is lacking. As discussed earlier, one method of validating novel non-experimental techniques is a WSC.

I am aware of only one WSC that compares treatment effect estimates from ML-generated counterfactuals with an RCT benchmark (Prest et al., 2023). In this study, the authors compare the average treatment effect (ATE) estimate from an RCT—the "benchmark" estimate—with those derived from two-way fixed effects regressions using either a NECG or ML-generated counterfactuals. They find that, when there is data only from treated households, XGBoost's counterfactual predictions are able to replicate the experimental benchmark, both when there is a true treatment effect and when there is no true effect. The authors highlight the use of rich data in ML, and explain that, because treatment was implemented in windows of three hours, they used hourly data by necessity, and could not aggregate to a weekly or monthly level to compare algorithm performance. Prest et al. (2023) provides some evidence of validity for the ML-generated artificial counterfactual method; however, it is unclear in which other contexts and data conditions this method is likely to be valid.

2.3 Contributions

While I also perform a WSC of ML artificial counterfactual methods, unlike Prest et al. (2023), I am concerned with water consumption rather than electricity usage. There are previous papers that use ML methods to predict water consumption (Walker et al., 2015; Shuang & Zhao, 2021; Kalashak, 2021; Dailisan et al., 2022; Kesornsit & Sirisathitkul, 2022), but most are at the city or region level rather than household, none predict counterfactuals in order to estimate a treatment effect, and certainly none perform WSCs. Also, unlike Prest et al. (2023), I use monthly observations rather than hourly. While highly granular data may often be available in energy, hourly data—or minutely, as the authors had before aggregation—is much more uncommon in other fields. Therefore, my main contributions are as follows: 1) I build on emerging ML artificial counterfactual prediction methods by using lower-frequency data to create predictions for monthly water use, and 2) I use a WSC framework to evaluate the ability of ML methods in these novel circumstances to uncover the benchmark treatment effect estimate derived from an RCT.

3. Data

I use data from a randomized experiment conducted by a research team from the Tropical Agricultural Research and Higher Education Center (CATIE), an academic center that studies development and environmental programs in Central America. 1,310 households across 9 communities in rural Costa Rica participated in the RCT, which was designed to measure the effect of water-efficient shower heads and faucet aerators on monthly water consumption. In addition to the RCT data, I use historical weather data from Visual Crossing, an online weather database.

3.1 RCT Benchmark

The CATIE RCT that I use as my benchmark was evaluated by Alpízar et al. (2023), and the treatment effect estimate using the full sample (both treated and control households), will serve as the experimental estimate by which I will evaluate the success of my ML-generated artificial counterfactual approach.

To obtain a sample of households for the experiment, CATIE researchers contacted communities whose water distribution systems are run by a community-based water management organization (CBWMO), and identified candidate ones according to three criteria:

1. Their CBWMO measured water use via meters and applied variable-rate pricing.
2. Their CBWMO had monthly household water use records dating back to 2012 and were willing to share them.
3. Their CBWMO would agree for the CATIE project team to install the water-efficient technologies randomly and were willing to share post-treatment monthly household water use data.

Of the 66 CBWMOs that CATIE contacted, 10 met these criteria, and 9 were selected to be used in the experiment. These 9 communities had a total of 2,246 customers, of which 1,898 were non-vacant, individually-metered residential properties. CATIE teams, each of which consisted of one interviewer and one plumber, approached all such households between May and July 2015, and were able to contact 1,346 of them. The

interviewer read a script that introduced the team members, informed the head of household about a CATIE climate study regarding local weather changes and their expected impact on water conservation, described and showed a video of the two water-efficient technologies, and offered to install the technologies for free if their home was randomly selected. 1,310 households agreed to have the CATIE team install the technologies same-day if they were selected. These 1,310 households comprise the experimental sample.

Households were randomized into one of three treatment arms depending on which color chip the resident selected at random out of an opaque bag. According to this procedure, 440 households were randomized into the control group and did not receive the technologies, 432 were randomized into the “no bonus" treatment group, and 438 were randomized into the “bonus" treatment group. Both households in the “no bonus” and “bonus” treatment groups received the technologies, but households in the “bonus” treatment group were offered a bonus of $38 USD if they still had all technologies installed at an unannounced follow up within the next 6 months. The bonus group is the focus of another study; for the purposes of this study I combine the two “no bonus” and “bonus” treatment arms into one treatment group.

3.2 ML Data

To recreate the constraints a researcher would face when using observational data in which all units were treated, I perform the entire ML approach using data only from the treated households of the RCT. These data include households’ socioeconomic and demographic information, home characteristics, community, as well as month and year of measurement. These variables, in addition to community-level weather ones, are used to train an ML model designed to predict monthly water consumption.[4]

Weather data were obtained from Visual Crossing, an online weather database that integrates data from local weather stations with NASA satellite and doppler radar data to provide a range of daily weather measures for most locations across the world

[4] While household characteristics are constant across the entire study period, the inclusion of monthly weather variables and month dummies allow for month-to-month variation in predictions.

(Visual Crossing, 2020). For each of the nine communities, daily data on maximum temperature, minimum temperature, mean temperature, precipitation, humidity, cloud cover, and UV index were extracted. These measures were aggregated to a monthly level and then merged with the RCT data on community. Summary statistics for both these weather variables and some of the household characteristics are presented in Table 1.

Table 1. Summary Statistics

Household Variables (N=870)	Treated	
	Mean	S.D.
Household size	3.6747	1.7831
Number of showers	1.0287	0.3001
Number of kitchen faucets	0.7724	0.4613
Number of bathroom faucets	0.5839	0.5810
At least one of the above fixtures	0.4391	0.4966
Attended primary school	0.8092	0.3932
Attended secondary school	0.2667	0.4425
Own home	0.8747	0.3312
Years in same home	18.2408	15.3103

Weather Variables (N=9)	Summer		Winter	
	Mean	S.D.	Mean	S.D.
Mean temperature (°F)	83.6694	1.1486	81.8612	1.3594
Max temperature (°F)	92.3597	4.0396	89.7096	4.4185
Minimum temperature (°F)	73.1236	6.8526	74.8169	4.6842
Total precipitation (in.)	0.7033	1.1384	9.1451	5.2761
Humidity (%)	64.8172	6.6084	78.8801	5.9458
Mean cloud cover (%)	46.0991	17.2133	72.3388	11.6226
Mean UV index	9.1677	0.5541	6.2870	0.8757

Table displays means and standard deviations for household characteristics and weather variables. The values for weather variables are across all nine communities for the associated months between January 2013 and September 2016. The dry "summer" season in Costa Rica is between January and April, while the wet "winter" season is between May and December.

4. Methods

In this section I detail my empirical approach. First, I explain the general structure of a supervised ML learning task and best practices when training an ML model. Next, I describe the supervised learning task as it applies to my study, as well as the specific ML algorithm used. I then explain how the fitted ML model is used to generate household-by-month counterfactual predictions, and how these predictions are used to estimate a treatment effect. Finally, I describe the criteria for a successful within-study comparison, assess the extent to which my study meets each criterion, and evaluate different measures of correspondence.

4.1 Machine Learning

Machine learning can be used to perform tasks in a variety of learning scenarios, including supervised learning, unsupervised learning, and semi-supervised learning[5] (Mohri et al., 2012). In supervised learning tasks the programmer provides the chosen algorithm with data on both inputs—called ***features*** (analogous to independent variables in standard regression models)—and the desired output—called the ***label*** (analogous to the dependent variable)—for a given number of observations or cases—called ***examples***. This series of labeled examples is known as the ***training set***, on which the algorithm fits a model by learning the combination of relationships between the features and labels that minimizes a certain ***loss function,***[6] which measures the difference between predicted and actual labels (Mohri et al., 2012). Given feature data for a series of unseen examples—known as the ***test set***—the fitted model can then be used to create a predicted label for each case. Quantitatively comparing the test set label predictions to the actual, unseen labels allows the programmer to evaluate the

[5] In unsupervised learning tasks all examples are unlabeled, and in semi-supervised learning tasks the learner receives a mix of both labeled and unlabeled examples (Mohri et al., 2012). Because I have monthly water consumption data available for all households, at least for a large majority of the sample period, my approach is considered a supervised learning task.

[6] Common loss functions are mean squared error (MSE or L2 loss) and mean absolute error (MAE or L1 loss).

model's performance via a ***score***[7].

A programmer can improve a model's performance via ***tuning***, during which they adjust the values for their chosen algorithm's ***hyperparameters***[8], which control the learning process. Tuning should be performed using a ***validation set***, a subset of the training data that is held out for evaluation during tuning, but is separate from the true holdout test data that is used to produce the final model score (Mohri et al., 2012). Since the validation set is also held out during model training, the validation score provides an *estimate* of the test score. Further, since it is separate from the true test set that is held out until tuning is completed, repeatedly refitting a model with different hyperparameters to maximize the validation score will not invalidate the model by biasing the test score. Rather, one should then be concerned about overfitting.

Using a single set of training and validation data to train and tune a model often leads to ***overfitting***, where an algorithm learns an overly complex model that performs extremely well on the particular data it has seen (i.e. the specific training and validation sets used for training and tuning), but poorly on unseen data (i.e. the holdout test set) (Mohri et al., 2012). A programmer can combat overfitting by using ***K-fold cross validation*** (CV). In *K*-fold CV the data is split into *K* subsamples, or "folds", where a single fold is held out as the validation set and used to calculate a model score while the other *K*-1 folds make up the training set. This process is repeated *K* times, typically without shuffling in between, until each fold has been used once as the validation set. An overall model score can be produced as an average of the score produced from each split (Pedregosa et al., 2011). *K*-fold CV helps to prevent overfitting by allowing a programmer to select a model with the best out-of-sample performance across *K* different unseen splits, rather than just one split.

I will now detail the supervised learning task as it is applied to my data, as well as the specific ML algorithm that is used for model training.

---

[7] The programmer must select the measure(s) that will be used to evaluate model performance. For regression problems, scoring methods include R^2, negative MSE, negative RMSE, negative mean absolute error, or negative median absolute error. A full list of the model scoring methods available via Python package scikit-learn can be viewed in the package's documentation on model selection.

[8] Hyperparameters and their purposes vary by algorithm. See section 4.1.a for a discussion of my chosen algorithm's hyperparameters.

4.1.a Model Training

Because my aim is to use a fitted ML model to generate predictions for monthly water use in the post-treatment time period for treated households *in the event that only treated data is available* and as if *no treatment had occurred*, I must train, tune, and test my model on only pre-treatment data from only the treated households. Using data from only the treated households recreates the constraints a researcher would face when using single-group treated-only data, and using exclusively pre-treatment examples to create the fitted model ensures that only the pre-treatment relationships between covariates and water use are learned.

Therefore, I train my ML model on only pre-treatment examples—between January 2013 and April 2015—from the treated households of the original RCT. To ensure that the algorithm sees data from every month and a representative sample of values for the label when training the model, I use Python package *scikit-learn* to perform a stratified split where, within each pre-treatment month, the training and test sets are balanced on bins of every 5th centile for average monthly household water. Following the stratified split, I proceed with training a model using ML algorithm XGBoost.

XGBoost stands for "Extreme Gradient Boosting" and is a gradient-boosted tree algorithm that can be used for both classification and regression problems. Classification and regression trees (CART) are grown by recursively partitioning data along the available inputs, where observations are sorted at each ***decision node*** into one of two ***branches*** according to their values for that input. Observations are split at the input variable and value that optimize the ***objective function***, a linear combination of the loss function and a regularization term that penalizes complexity to avoid overfitting. The predicted outputs are obtained from the sample average of observations in each terminal node, or ***leaf***. Instead of growing only a single classification or regression tree, XGBoost grows an ensemble of trees sequentially, where each successive tree is estimated on the residual of the previous tree to correct that tree's errors—this is called ***boosting***. The weight of each new tree is scaled between boosting rounds by a factor η to reduce the importance of any one tree—this

is called ***shrinkage***. The final XGBoost model is the aggregation of the weighted trees. (Chen & Guestrin, 2016)

Table 2. Hyperparameter Tuning

Hyperparameter	Value
Number of trees (T)	100
Maximum tree depth	50
Shrinkage (η)	0.13
Minimum loss reduction for split (γ)	10
L1 regularization on weights (α)	10
L2 regularization on weights (λ)	20
Fraction of training sample used per tree	0.75
Fraction of features sampled per tree	0.75

I implement XGBoost using Python package *xgboost*. There are a variety of hyperparameters that can be chosen to influence XGBoost's learning process and control the resulting model's complexity[9]. I select values for the hyperparameters shown in Table 2 by tuning via 5-fold CV. Higher values for the number of trees and maximum tree depth result in a more complex model, while higher values for shrinkage, minimum loss reduction, the two regularization terms, and the fractions of training examples and features sampled per tree control overfitting. All other hyperparameters were left at their default values.

4.1.b Generating Counterfactuals

After training and tuning my ML model on the pre-treatment, treated-only data, I use the fitted model to generate predictions for water consumption for each household

[9] A full list of the available hyperparameters, their descriptions, and default values is available in the XGBoost Python package documentation.

for each month. These predictions are denoted $\hat{Y}_{it}$. The post-treatment (June/July[10] through September 2016) predictions serve as counterfactuals—hypothetical values for what households' water consumption would have been if treatment had not occurred—because they are calculated according to the pre-treatment relationships between the features and water use. Absent treatment, and ceteris paribus, it is reasonable to assume these relationships would have persisted. The mean monthly household actual and ML-predicted water consumption is shown below.

Figure 1. Actual vs. Predicted Water Use

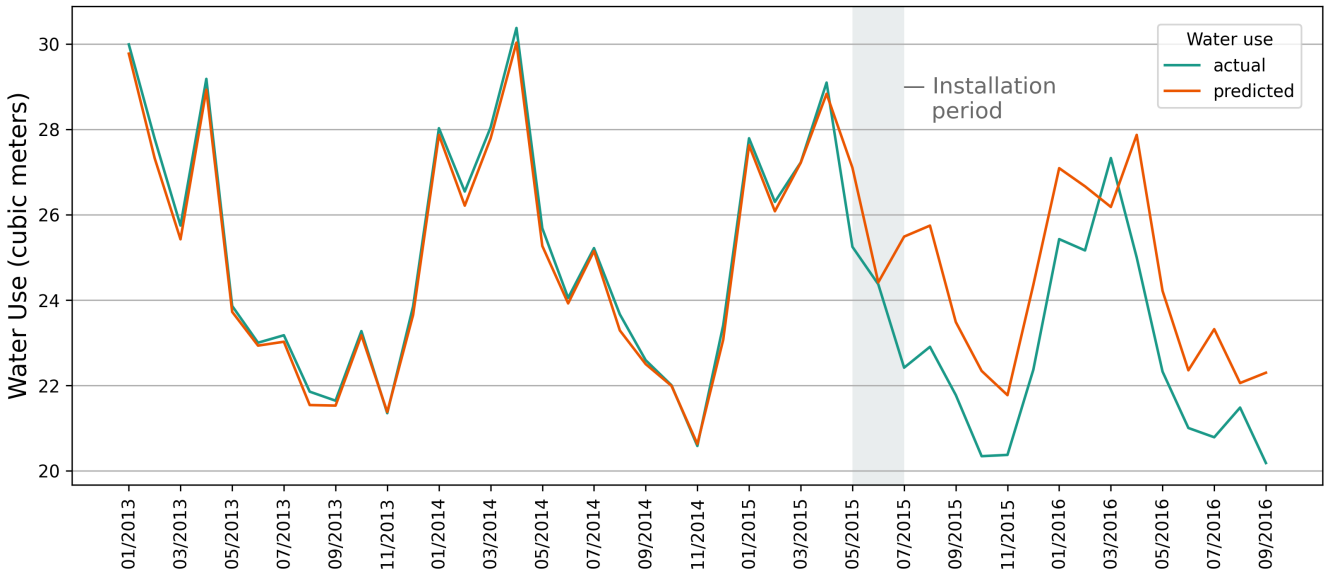

Figure displays mean actual and ML-predicted household water use (in cubic meters) for each month. The period during which the water-efficient technologies were installed is shaded in gray.

4.2 Treatment Effect Estimation

As in Alpízar et al. (2023), I seek to calculate the ATE of water-efficient technology installation on monthly household water use over the 16-month post-

[10] The technologies were installed in treated households in either May or June (or the first week of July; these households are considered to be treated in June). For this reason, only observations though April 2015 were used for model training. Predictions, however, were generated for all months; only those for households' post-installation months represent post-treatment counterfactuals.

treatment period.[11] I begin by calculating the household-by-month prediction error as the difference between actual and predicted consumption, $Y_{it} - \hat{Y}_{it}$. If the water-efficient shower heads and faucet aerators did in fact reduce households' monthly water consumption, one would expect the actual consumption to be lower than the predicted in the post-treatment period, and, consequently, the prediction error to be negative. While the average prediction error across all households and post-treatment months would give some indication of the ATE, it does not properly account for the similarities that may exist in prediction errors within a given household or month.

As a result, I use the fixed effects model:

$$Y_{it} - \hat{Y}_{it} = \beta_0 + \beta_1 \cdot post\_treat_{it} + train_{it} + household_i + month_j + \mu_{it} \qquad (1)$$

where $post\_treat_{it}$ is a dummy equal to zero in pre-treatment months and one in post-treatment months, and where $household_i$ is a dummy for each household. The household fixed effects will capture the extent to which the ML prediction model systematically under- or over-estimates water use within each household, and the month dummies will capture seasonal patterns, or the extent to which the model under- or over-estimates water use during a given month of the year. The dummy variable $train_{it}$ is equal to one for observations that appeared in the training set and zero for observations that did not. This dummy is included in order to account for the fact that, since the model is created in order to optimize performance on the training set, the prediction error of observations in the training set is expected to be smaller than the prediction error of observations on which the model was not trained. In order to match Alpízar et al. (2023), this regression is run using data only from May 2014 on, with the months January 2013 through April 2014 excluded.

Because the dependent variable in my fixed effects model is the prediction error, or difference between actual and predicted water use, the estimate for β_1, the coefficient of $post\_treat_{it}$, will indicate the amount by which actual water use was reduced in the post-treatment period relative to the counterfactuals, or the values for water use that would have been expected absent treatment assuming the pre-

[11] Given potential disadoption at later dates, this estimand is not the same as the ATE of adopting and keeping the technologies installed for the entire post-treatment period.

treatment trends that the ML algorithm learned would have persisted. Therefore, $\hat{\beta}_1$ represents my quasi-experimental treatment effect estimate.

Because error is introduced in the estimate of β_1 during both steps of the two-step estimation strategy—first in the ML counterfactual prediction process and then again in the fixed effects regression—and there is no way to carry through the error of the first step, I perform a cluster bootstrap procedure to estimate the standard error for $\hat{\beta}_1$, clustered at the household-level. Within each repetition, the original panel data is resampled with replacement by cluster, meaning that for a household selected n times during resampling, and with m months of data recorded, there will be $n \cdot m$ observations in the resampled data that are associated with that household. The ML counterfactual prediction method is then applied to these resampled data (save for tuning, which would be computationally infeasible to repeat each round), and the resulting predictions used in the previously-specified fixed effects regression to obtain an estimate of β_1 specific to that bootstrap repetition, denoted θ. This process was repeated—first resampling, then ML counterfactual prediction, then treatment effect estimation via the fixed effects regression—for 250 iterations. The bootstrapped standard error, $SE(\hat{\beta}_1)$, is equal to the standard deviation of the bootstrap repetition-specific coefficient estimates:

$$SE(\hat{\beta}_1) = \sqrt{\sum_{k=1}^{250} \frac{(\theta_k - \bar{\theta})^2}{250 - 1}} \qquad (2)$$

To ensure a fair comparison between the experimental and quasi-experimental estimates, the benchmark ATE was re-estimated using a fixed effects regression of the form:

$$Y_{it} = \alpha_0 + \alpha_1 \cdot post_{it} + \alpha_2 \cdot post\_treat_{it} + household_i + month_j + \mu_{it} \qquad (3)$$

where $post\_treat_{it}$ is equal to one for observations corresponding to treated households in post-treatment months and zero otherwise. In this model, $\hat{\alpha}_2$ is the

treatment effect estimate, and this is the value against which I will compare my quasi-experimental estimate, $\hat{\beta}_1$, using a within-study comparison framework.

4.3 Within-Study Comparison

Dependent-arm within-study comparisons[12] compare the results of an RCT to the results of a quasi-experiment that shares some portion, typically the treatment group, of the original experimental sample (Wong & Steiner, 2018). In order to be sure that one is making a valid comparison, and that any detected difference in the obtained estimates stems only from a failure of the quasi-experiment, rather than a poor WSC design, researchers have developed guidelines that, when followed, result in a credible WSC.

4.3.a Criteria for a Successful Within-Study Comparison

There exist several criteria for designing a successful WSC. The criteria put forth by Cook et al. (2008) are as follows: (1) the WSC includes both a randomly-assigned counterfactual group (i.e. the control group) and a nonrandom one, (2) the experiment and quasi- or non-experiment estimate the same causal quantity (e.g. ATE, ITT, etc.), (3) the selection of the experimental sample and the quasi-experimental sample should not be correlated with other variables that are related to the outcome of interest, (4) analysts of the experiment and quasi- or non-experiment should be blind to each other's results, (5), the experiment should meet the usual criteria for technical adequacy (e.g. proper randomization, low non-compliance), (6) the quasi or non-experiment should meet the standard criteria for technical adequacy, and (7) some measure(s) of correspondence is adopted to compare the causal quantities estimated by the experiment and the quasi- or non-experiment.

In my study context I am able to meet all but one of the criteria from Cook et al. (2008). My study does not compare a randomly-assigned counterfactual group to a nonrandom comparison group, but instead a randomly-assigned counterfactual group to an artificial one created from the original randomized treatment group (1). The

[12] These are in contrast to independent-arm WSCs, in which units are randomly assigned either to the RCT or quasi-experimental arm, and no portion of the sample is shared (Wong & Steiner, 2018).

experimental and quasi-experimental analyses both estimate ATE effects (2). Selection of the experimental and quasi-experimental samples is not correlated with any variables that are related to the outcome (3). I was blind to the results of the original experiment while developing my quasi-experimental estimate (4). The experiment and its analysis were conducted according to field standards (5). The ML-generated artificial counterfactual predictions were obtained according to best practices in ML as described in Section 4.1 (6). My discussion of the final criterion follows in the selection below.

4.3.b Measures of Correspondence

As noted by Cook et al. (2008) and Steiner & Wong (2018), there exists no field-wide standard for evaluating the similarity of the treatment effect estimate obtained from the RCT, denoted T_E, and from the quasi-experiment, denoted T_{NE}. Some common correspondence criteria that have been used in the WSC literature include: 1) T_{NE} falls within a certain number of standard deviations from T_E, 2) the point estimate for T_{NE} falls inside a certain confidence interval of T_E, or 3) the confidence intervals of T_{NE} and T_E overlap. Two common values used for the standard deviation criteria are 0.1 (Cook et al., 2020; Coopersmith et al., 2022) or 0.2 (St. Clair et al., 2014; St. Clair et al., 2016) standard deviations. Confidence intervals, when used, are typically 95% confidence intervals (Fretheim et al., 2013; Fretheim et al., 2015; Ferraro & Miranda, 2014; Prest et al., 2023), except in the case of Shadish et al. (2016), who use overlapping 84% confidence intervals as their correspondence criteria to reduce the possibility of making a Type II error (i.e. finding correspondence when there is not).

To create a more standardized measure, Steiner & Wong (2018) suggest a ***correspondence test***, for which equivalence of the estimates is found only if there is both a significant equivalence, C_E, and an insignificant difference, C_D.

Table 3. Steiner & Wong (2018) Correspondence Test

		Significant equivalence (C_E)	
		Yes ($C_E = 1$)	No ($C_E = 0$)
Insignificant difference (C_D)	Yes ($C_D = 1$)	equivalence	indeterminacy
	No ($C_D = 0$)	trivial difference	difference

Significant equivalence (C_E) is achieved if one can reject the null hypothesis that the absolute difference $|T_E - T_{NE}|$ is statistically significantly greater than δ, a pre-defined threshold chosen by the researcher (e.g. $\delta = 0.1\ SD$). This entails a one sided t-test using the t-statistic $t = \frac{|T_{NE} - T_E|}{s}$, where s is equal to the standard error of the effect difference as obtained via bootstrap[13]. An **insignificant difference** (C_D) is achieved if one fails to reject the null hypothesis $|T_E - T_{NE}| = 0$ using the same t-statistic as the equivalence test. This two-part correspondence test is regarded to be more rigorous because underpowered WSCs, which may have concluded correspondence given only C_D, will instead result in an indeterminate outcome if the tolerance threshold for C_E is appropriately low. Of the five WSCs I was able to identify that used the the Steiner & Wong (2018) correspondence (besides those authored by the creators), all found indeterminacy, trivial difference, or difference in the majority of the correspondence tests they performed (Altindag et al., 2019; Anderson & Wolf, 2019; Litwok, 2020; Anderson et al., 2021; Unlu et al., 2021).

To evaluate the success of my ML counterfactual prediction method, I decide to adopt the correspondence measure proposed by Steiner & Wong (2018). I select this measure because: 1) it is stringent and should not result in an erroneous conclusion of correspondence, and 2) in combining the two separate C_E and C_D tests and allowing

[13] Steiner & Wong (2018) recommend bootstrapping to obtain the standard error of the effect difference due to dependency between the experimental and quasi-experimental samples (since the treatment group is shared by both). I perform 250 bootstrap iterations for estimation of this standard error.

for an indeterminate result, it enables one to perform a more nuanced assessment of correspondence. I use a threshold of $\delta = 0.1\ SD$ to determine C_E, and $\alpha = 0.05$ as the significance level for both C_E and C_D. In addition to the Steiner & Wong (2018) correspondence test, for sensitivity analysis, I present the results according to the second and third measures described in the first paragraph of this section[14].

5. Results

In this section I present the results of the ML counterfactual prediction method and within-study comparison. I begin with a table summarizing the fixed effects prediction error regression results. After discussing the results shown in the table, I present the outcomes of the correspondence tests described in Section 4.3.b. Finally, to determine how much—if any—benefit is had from using ML over a simpler single-group design such as counterfactual prediction method using OLS or single interrupted time series (SITS), I compare the results of these other methods to those of the ML counterfactual prediction approach.

5.1 Fixed Effects Regression Estimates

The results of the randomized experiment and the results of the ML approach using the fixed effects prediction error regression described in Section 4.2 are shown in Table 4. The first column shows the results of the experiment as estimated using Equation 3 in Section 4.2, and the second column shows the results of the ML quasi-experiment as estimated using Equation 1 in the same section. The coefficient of $post_{it}$ in the first column can be interpreted as the amount that monthly household water use changed from the pre- to post-treatment period for the control households, on average. This estimate's statistical insignificance provides evidence against the occurrence of history, or an event concurrent with but separate from treatment that also affected household water use, and gives support to my belief that it is reasonable to assume that the pre-treatment trends persist into the post-treatment period. Further

[14] The first measure (the quasi-experimental treatment effect estimate falls within a certain number of standard deviations from the experimental benchmark) is already embedded in the Steiner & Wong (2018) correspondence test that I select as my primary measure of similarity.

support of this conclusion can be obtained from an exercise using data from only the control group rather than the treated, during which the ML counterfactual prediction method and treatment effect estimation approach were applied to these data instead. This exercise is effectively a placebo test and in fact failed to find evidence of a statistically significant treatment effect for the control households[15].

Table 4. Fixed Effect Regression Results

	Water use (cubic meters)	
	RCT	ML
post	0.10324	
	(0.3564)	
post_treat	-2.1273***	-1.7327***
	(0.4410)	(0.3325)
train		0.1139
		(0.1913)
constant	27.7742***	0.0668
	(0.2189)	(0.3695)
number of observations	37509	24894
number of households	1309	870

Table displays regression coefficients for estimates of monthly household water use, measured in cubic meters. Cluster-robust standard errors are in parentheses. Standard errors for the second column are bootstrapped according to the procedure described in Section 4.2.

*p<0.1, **p<.05, ***p<0.01

[15] Results of the control group exercise are shown in appendix table A.1.

The benchmark ATE from the RCT is given by the coefficient of $post\_treat_{it}$ in the first column, equal to a 2.13 cubic meter reduction in monthly household water use, on average. The asterisks indicate statistical significance of the estimate. The comparable ML quasi-experimental estimate is the coefficient of the same variable in the second column. As seen in the table, the ML approach yielded an estimate of an average reduction in monthly household water consumption equal to approximately 1.73 cubic meters. While the point estimate given by the ML approach is the correct sign, true correspondence is not achieved unless the estimates are equivalent according to my selected correspondence test.

5.2 Correspondence Testing

I present first the results of my primary measure of correspondence, the Steiner & Wong (2018) correspondence test. The test statistics as described in Section 4.3.b as well as the associated levels of statistical significance and conclusions are shown in Table 5. As seen in the table, the test found both significant equivalence and an insignificant difference, which, together, result in a conclusion of equivalence between the two estimates.

Table 5. Correspondence Test Results

	T-statistic	Conclusion
Significant equivalence (C_E)	-2.9891***	Yes ($C_E = 1$)
Insignificant difference (C_D)	0.9575	Yes ($C_D = 1$)
Result		equivalence

Table displays results of Steiner & Wong (2018) correspondence test. T-statistics were produced using bootstrapped standard errors as described in Section 4.3.b.

*p<0.1, **p<.05, ***p<0.01

Steiner & Wong (2018) acknowledge that selection of the tolerance threshold δ is discretionary, and while I did select values commonly used in the WSC literature, I do

admit they were otherwise arbitrary. For this reason, and following the suggestion of Steiner & Wong (2018), I report the smallest threshold for which equivalence would still be concluded, which is equal to $0.0661\ SD$[16].

As one may expect given the results of this correspondence test, the ATE estimates are also considered equivalent according to the less rigorous measures described at the beginning of Section 4.1.b: the ML quasi-experimental point estimate falls inside the 95% confidence interval of the experimental estimate, and, consequently, their 95% confidence intervals overlap (Figure 2).

Figure 2. RCT Benchmark and ML Approach Treatment Effect and CIs

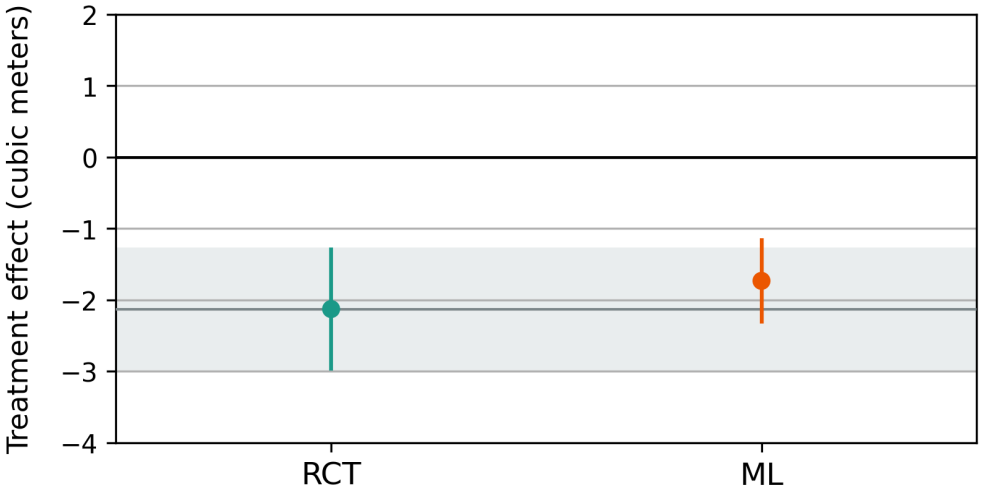

Figure displays point estimates and 95% confidence intervals for the benchmark ATE and the ATE as estimated by the ML approach. The 95% confidence for the benchmark is shaded.

5.3 Comparison to Alternative Single-Group Designs

In order to determine if there is any advantage gained from using machine learning over a simpler approach, I present the ATE as estimated via two alternative single-group quasi-experimental designs (Table 6): a non-ML counterfactual prediction

[16] Of note, the threshold is defined in terms of standard deviations of the outcome (water use, in this context) for the control group. For the period May 2014 through September 2016 (the same period used for ATE estimation), the standard deviation of monthly household water consumption for the control group was 16.26 cubic meters, which, when considering an effect size of 2.12 cubic meters, is comparatively quite large.

method that uses linear regression for the prediction model, and single interrupted time series.

Table 6. Benchmark vs. Single-Group Quasi-Experimental ATE Estimates

	Water use (cubic meters)			
	RCT	ML	Non-ML	SITS
ATE estimate	-2.1273***	-1.7327***	-2.3340***	-1.8117***
	(0.4410)	(0.3325)	(0.2862)	(0.4643)
number of observations	37509	24894	22993	35338
number of households	1309	870	803	803

Table displays estimated treatment effects for each method, measured in cubic meters. Cluster-robust standard errors are in parentheses. Standard errors for the second and third columns are bootstrapped according to the procedure described in Section 4.2.

*p<0.1, **p<.05, ***p<0.01

The non-ML method uses the same prediction error approach as the ML method in order to estimate the ATE, but uses an OLS model fitted on pre-treatment data rather than XGBoost to generate the predictions for household-by-month water consumption. The full list of variables used in the OLS model is shown in appendix table A.2. Once the predictions are calculated according to this OLS model, the fixed effects prediction error regression (appendix equation A.1) is used to estimate the ATE[17]. As with the ML approach, and using the same procedure, the standard error is estimated via bootstrapping. As seen in Table 6, the non-ML approach yields an estimated ATE of -2.33 cubic meters.

I produce the single interrupted time series estimate according to the model described in Huitema & McKean (2000)[18], with the error term modeled by an AR(1)

[17] This regression is the same as Equation 1, barring exclusion of the train dummy. In the non-ML approach, all pre-treatment data was used to fit the OLS model; therefore, the train dummy would be perfectly collinear with the post-treatment dummy.

process[19] (appendix equation A.2). Unlike the ML and non-ML approaches, which involve two steps—a prediction step (using either XGBoost or OLS) and an estimation step (using a fixed effects regression)—SITS involves only one[20]. The SITS model estimates a time trend for the entire period, an immediate effect (i.e. intercept change) upon treatment, and allows for a new time trend post-treatment. I include additional controls for household characteristics (number of people, income indicators, and number of devices/appliances that use water), community, interactions between community and interviewer teams[21], weather, as well as month to capture seasonality. I implement SITS using the *prais* command in Stata, which "uses the generalized-least squares method to estimate the parameters in a linear regression model in which the errors are serially correlated […, and] are assumed to follow a first-order autoregressive process." (Hardin & StataCorp, n.d.). SITS produces an estimated ATE[22] of approximately -1.81 cubic meters.

All three of the single-group quasi-experimental designs are able to produce ATE estimates with the correct sign. The non-ML counterfactual prediction method using OLS produces the closest point estimate to the experimental one, with an absolute difference of approximately 0.2 cubic meters. SITS produces the next-closest estimate, and the ML counterfactual prediction method produces the furthest estimate. All three point estimates fall within the 95% confidence interval of the benchmark (Figure 3).

---

[19] An AR(1) process is used to model the error term as the Durbin-Watson test of the unadjusted errors returns a test statistic of 0.47, indicating autocorrelation. I consider only one lag because the Durbin-Watson test returns a statistic of approximately 2.3 for the transformed errors, failing to find evidence of further autocorrelation beyond one lag.

[20] For this reason, the SITS regression was run for all months (January 2013 through September 2016) rather than the May 2014 though September 2016 period used for the fixed effects regression applied to the ML and non-ML methods.

[21] This interaction is included because each interviewer team visited a cluster of households within a neighborhood; therefore, households within the same cluster that are visited by the same team may be similar in ways that are not fully accounted for by the available control variables.

[22] Because the SITS model decomposes the treatment effect into immediate and sustained effects, the reported estimate for the ATE for the 16 month post-treatment period must be calculated—in order to produce a comparable estimate—using both effects. Stata command *lincom* was used to produce a standard error for the treatment effect estimate, which is a linear combination of three coefficients in the SITS model.

Figure 3. Benchmark vs. Single-Group Quasi-Experimental ATE Estimates

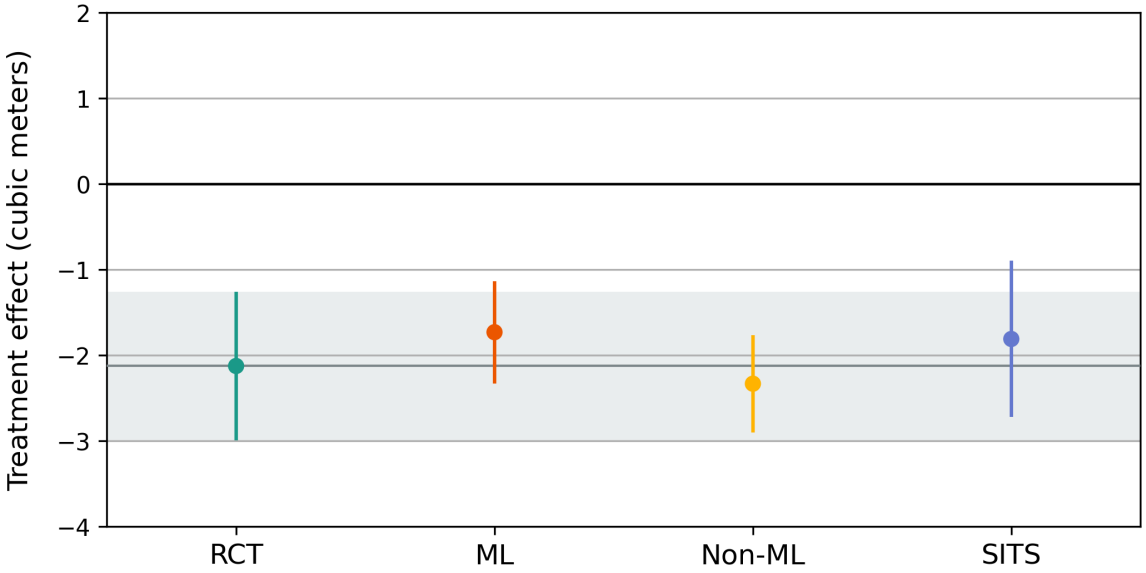

Figure displays point estimates and 95% confidence intervals for the ATE as estimated by each approach. The 95% confidence for the benchmark is shaded.

Because it is difficult to compare the potential benefit gained from using ML over one of the aforementioned alternative approaches when the primary measure of a quasi-experimental design's usefulness is its ability to replicate experimental findings (and therefore demonstrate equivalence using a correspondence test), correspondence testing was also performed for both of the alternative methods. For each design, I report the conclusion of the Steiner & Wong (2018) correspondence test with $\delta = 0.1\ SD$ and $\alpha = 0.05$ as well as the minimum threshold for which equivalence would be concluded using the same significance level (Table 7).

Table 7. Correspondence Testing of Quasi-Experimental ATE Estimates

	Conclusion	Min. Threshold
ML	equivalence	0.0661 SD
Non-ML	equivalence	0.0488 SD
SITS	equivalence	0.0518 SD

Table displays conclusion of Steiner & Wong (2018) correspondence test for each single-group quasi-experimental design using a threshold of $\delta = 0.1\ SD$ and significance level $\alpha = 0.05$, as well as the minimum threshold δ for which equivalence would be found.

I find that, of the three, the non-ML counterfactual prediction method using OLS would conclude equivalence at the smallest threshold, and, if I had selected a threshold of $0.05\ SD$, only this approach would result in a conclusion of equivalence. While this was initially a surprising result considering that ML models such as XGBoost are often regarded to be more powerful than simple OLS due to support for nonlinearities between features, it underscores the lesson that more complex models are not best-suited for all problems. Excepting the—again arbitrary—hypothetical alternative threshold of $0.05\ SD$, the minimum thresholds are otherwise quite similar across approaches.

6. Discussion

My primary results do provide evidence that the ML counterfactual prediction method can replicate experimental treatment effects outside of the context of Prest et al. (2023), in which the authors had access to highly granular energy data with a large number of observations. Even with only monthly data and a fraction of the observations used in their study, I find that the ML method is able to produce a treatment effect in the same direction as the original experimental one, and is "equivalent" according to a range of commonly-used correspondence criteria. This

further supports the validity of the ML counterfactual prediction method in estimating causal effects when data are available only from treated units, and suggests that the ML counterfactual prediction method may have value in situations even when observations number in the tens of thousands rather than millions.

However, in situations where this is true, it may also be the case that a simpler counterfactual prediction model may provide similar results, or, as in my case, offer a slight improvement over the more complex one. When I apply the same prediction error approach, but use an OLS regression rather than XGBoost to create the predictions, I find that the resulting ATE estimate is 0.19 cubic meters closer to the benchmark estimate than the one that was calculated using the XGBoost-generated predictions. Correspondence testing finds that the non-ML approach using OLS would be able to produce a conclusion of equivalence at a similar, though slightly lower tolerance threshold than the ML approach. Therefore, it could be the case that a simpler model is more appropriate for situations with a similar number of observations, as it may be difficult with a dataset of this size to identify complex, nonlinear interactions between variables that are generalizable to a larger population or longer period. If so, an intricate ML model like XGBoost would not likely offer any advantage over simple OLS, and a researcher may benefit from selecting a non-ML counterfactual prediction method instead.

While I find that all three of the single-group designs I tested were able to reproduce experimental results in this context, it is important to discuss their limitations. As previously noted, single-group designs are vulnerable to history, an event (or events) that occur(s) concurrently with treatment, but separate from treatment, that can affect the outcome of interest. Single-group quasi-experiments, by design, lack a control or comparison group that did not receive treatment but would still be affected by history, and are consequently unable to resolve this issue. Because I had access to the full RCT panel, including data from the control group, I was able to test for the existence of history, but a researcher interested in the practical application of one of these single-group designs, who has access to data from treated units only, would not. In such a situation, an understanding of the historical context surrounding the study period is critical. If the researcher believes that no history occurred, they

should be prepared to justify their reasoning. In the *presence* of history, researchers should be able to explain the mechanism(s) through which history may have affected the outcome variable, as well as the direction in which the treatment effect estimate would be biased as a result.

For example, if there had been a drought in Costa Rica that coincided with installation of the water-efficient technologies, two possible mechanisms for its impact on water use are: 1) households may be more conscious of their water consumption during a drought, and hence use less water, or 2) households may need to use extra household water to supplement use that would typically be supplied by rainfall (e.g. gardening). A better cultural understanding could help one to determine which of these mechanisms is more likely to apply. Assuming the former, a single-group quasi-experimental design like pretest-posttest would likely overestimate the ATE, as it would attribute the calculated reduction in water use following treatment only to the installation of the water-efficient technologies, when, in actuality, it also stemmed from households' changed pattern of consumption as a result of the drought. While it is possible that the counterfactual prediction method, with rainfall as an input, would accordingly predict reduced water use post-treatment (i.e. during the drought), without an event of similar magnitude in the training period, it is unlikely that it would be able to capture the full extent to which water use should be reduced as a result of the drought.

Therefore, although counterfactual prediction may offer a *slight* improvement in the presence of such an event, it is still important that a researcher be able to reason through such a mechanism and state the expected direction of bias. Because no such event was present in my data, I am unable to offer any insight about how much improvement there may be. Future research regarding the performance of counterfactual prediction methods in the presence of history could provide more information about the cases in which they are likely to offer improvements and those in which they are not.

7. Conclusion

RCTs are sometimes unfeasible to implement in practice, and appropriate comparison groups sometimes do not exist. In such cases, researchers are left with

data from treated units only, and must rely on single-group quasi-experimental designs in order to estimate program effects. This work adds to a small body of within-study comparison literature evaluating said single-group designs.

Although my primary ML-based counterfactual prediction method is able to reproduce the experimental ATE estimate, I find that a non-ML, OLS-based approach and a single interrupted time series design are able to as well, and—in fact—that the non-ML approach slightly outperforms the ML approach. These results can provide insight to future researchers regarding implementation of the ML counterfactual prediction method as well as the data conditions in which one may instead choose to opt for a simpler design, such as the non-ML counterfactual prediction method using a linear regression model. While this work suggests that ML may be not be the most suitable choice in situations with a dataset of this size, future research into the use of these single-group designs in similar data contexts is needed to further validate this conclusion.

8. References

Alpízar, F., Bernedo del Carpio, M., & Ferraro, P. (2023). Input Efficiency as a Solution to Externalities: A Randomized Controlled Trial. Working paper, Wageningen University and Research.

Altindag, O., Joyce, T. J., & Reeder, J. A.. (2019). Can Nonexperimental Methods Provide Unbiased Estimates of a Breastfeeding Intervention? A Within-Study Comparison of Peer Counseling in Oregon. *Evaluation Review*, 43(3-4), 152-188. https://doi.org/10.1177/0193841X19865963

Ambroggio, L., Smith, M. J., & Shah, S. S. (2021). Quasi-Experimental and Interrupted Time-Series Design [Editorial commentary of "Impact of a prospective-audit-with-feedback antimicrobial stewardship program at a children's hospital" by Newland, J. G., Stach, L. M, DeLurgio, S. A., Hedican, E., Yu, D., Herigon, J. C. , Prasad, P. A., Jackson, M. A., Myers, A. L., & Zaoutis, T.E.]. *Journal of the Pediatric Infectious Diseases Society,* 1(3), 187–189. https://doi.org/10.1093/jpids/pis059

Amiri, S. S., Mostafavi, M., Lee, E. R., & Hoque, S. (2020). Machine learning approaches for predicting household transportation energy use. *City and Environment Interactions*, 7. http://dx.doi.org/10.1016/j.cacint.2020.100044

Anderson, K. P. & Wolf, P. J. (2019). Does Method Matter? Assessing the Correspondence between Experimental and Nonexperimental Results from a School Voucher Program Evaluation. Working paper 2017-10, University of Arkansas, Department of Education Reform. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2952967

Anderson, K., Zamarro, G., Steele, J., & Miller, T. (2021). Comparing Performance of Methods to Deal With Differential Attrition in Randomized Experimental Evaluations. *Evaluation Review*, 45(1-2), 70–104. https://doi.org/10.1177/0193841X211034363

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2021). Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association,* 116 (536). https://doi.org/10.1080/01621459.2021.1891924

Ayuya, C. (2021). Minimizing Data Leakage in Machine Learning. *Section*. https://www.section.io/engineering-education/data-leakage/

Baicker, K. & Svoronos, T. (2019). Testing the Validity of the Single Interrupted Time Series Design. Working paper, National Bureau of Economic Research. http://www.nber.org/papers/w26080

Burlig, F., Knittel, C., Rapson, D., Reguant, M., & Wolfram, C. (2021). Machine Learning From Schools About Energy Efficiency. *Journal of the Association of Environmental and Research Economists*, 7(6), 1005-1217. https://doi.org/10.1086/710606

Chen, J. (2021). An Introduction to Machine Learning for Panel Data. *International Advances in Economic Research*, 27, 1-16. https://doi.org/10.1007/s11294-021-09815-6

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Association for Computing Machinery*. https://doi.org/10.1145/2939672.2939785

Christiansen, P., Francisco, P., Myers, E., & Souza, M. (2021). Decomposing the Wedge Between Projected and Realized Returns in Energy Efficiency Programs. *The Review of Economics and Statistics*. Forthcoming. https://doi.org/10.1162/rest_a_01087

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons. *Journal of Policy Analysis and Management*, 27(4), 724-750. https://doi.org/10.1002/pam.20375

Cook, T. D., Zhu, N., Klein, A., Starkey, P., & Thomas, J. (2020). How Much Bias Results if a Quasi-Experimental Design Combines Local Comparison Groups, a Pretest Outcome Measure and Other Covariates?: A Within Study Comparison of Preschool Effects. *Pyschological Methods*, 25(6), 726-746. http://dx.doi.org/10.1037/met0000260

Coopersmith, J., Cook, T. D., Zurovac, J., Chaplin, D., & Forrow, L. V. (2022). Internal and External Validity of the Comparative Interrupted Time-Series Design: A Meta-Analysis. *Journal of Policy Analysis and Management*, 41(1), 252-277. https://doi.org/10.1002/pam.22361

Dailisan, D., Liponhay, M., Alis, C., Monterola, C. (2022). Amenity counts significantly improve water consumption predictions. *PLOS One*, 17(3). https://doi.org/10.1371/journal.pone.0265771

Dueñas, M., Ortiz, V., Riccaboni, M., & Serti, F. (2021). Assessing the Impact of COVID-19 on Trade: a Machine Learning Counterfactual Analysis. Working paper 79, Red Investigadores de Economía. https://doi.org/10.48550/arXiv.2104.04570

Ferraro, P. J, & Miranda, J. J. (2014). Panel Data Designs and Estimators as Substitutes for Randomized Controlled Trials in the Evaluation of Public Programs. *Journal of the Association of Environmental and Resource Economists*, 4(1), 281-317. https://www.journals.uchicago.edu/doi/full/10.1086/689868

Fretheim, A., Soumerai, S. B., Zhang, F., Oxman, A. D., Ross-Degnan, D. (2013). Interrupted time-series analysis yielded an effect estimate concordant with the cluster-randomized controlled trial result. *Journal of Clinical Epidemiology*, 66(8), 883-887. https://doi.org/10.1016/j.jclinepi.2013.03.016

Fretheim, A., Zhang, F., Ross-Degnan, D., Oxman, A. D., Cheyne, H., Foy, R., Goodacre, S., Herrin, J., Kerse, N., McKinlay, R. J., Wright, A., Soumerai, S. B. (2015). A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation. *Journal of Clinical Epidemiology*, 68(3), 324-333. https://doi.org/10.1016/j.jclinepi.2014.10.003

Hardin, J. W. & StataCorp. (n.d.) prais: Prais–Winsten and Cochrane–Orcutt regression. *StataCorp*. https://www.stata.com/manuals/tsprais.pdf

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2016). Counterfactual Prediction with Deep Instrumental Variables Networks. https://doi.org/10.48550/arXiv.1612.09596

Huitema, B. & McKean, J. (2000). Design Specification Issues in Time-Series Intervention Models, Educational and Psychological Measurement, 60(1), 38-58. https://doi.org/10.1177/00131640021970358

Kalashak, E. (2021). Prediction of Water Consumption Using Machine Learning: Using machine learning techniques to predict hourly water consumption in sustainable smart city. Master's thesis, Østfold University College. https://hiof.brage.unit.no/hiof-xmlui/bitstream/handle/11250/2778632/Kalashak_Elahe.PDF?sequence=1&isAllowed=y

Kesornsit, W. & Sirisathitkul, Y. (2022). Water consumption prediction based on machine learning methods and public data. *Advances in Computational Design*, 7(2), 113-128. https://doi.org/10.12989/acd.2022.7.2.113

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Association*, 76(4), 604-620. https://www.jstor.org/stable/1806062

Litwok, D. (2020). Using Nonexperimental Methods to Address Noncompliance. Working paper 20-324, *Upjohn Institute for Employment Research*. https://doi.org/10.17848/wp20-324

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2012). *Foundations of Machine Learning.* MIT Press. Print.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, 12(85), 2825–2830. https://scikit-learn.org/stable/modules/grid_search.html#grid-search-tips, https://scikit-learn.org/stable/modules/grid_search.html#grid-search-tips

Prest, B., Wichman, C., & Palmer, K. (2023). RCTs against the machine: Can machine learning prediction methods recover experimental treatment effects? J*ournal of the Association of Environmental ad Resource Economists,* forthcoming. https://doi.org/10.1086/724518

Raschka, S. & Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. (3rd ed.). Packt.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company. Print.

Shadish, W. R., Rindskopf, D. M, & Boyajian, J. G. (2016). Single-case experimental design yielded an effect estimate corresponding to a randomized controlled trial. *Journal of Clinical Epidemiology*, 76, 82-88. https://doi.org/10.1016/j.jclinepi.2016.01.035

Shuang, Q., Zhao, R. T. (2021). Water Demand Prediction Using Machine Learning Methods: A Case Study of the Beijing–Tianjin–Hebei Region in China. *Water*, 13, 310. https://doi.org/10.3390/w13030310

Souza, M. (2022). Predictive Counterfactuals for Treatment Effect Heterogeneity in Event Studies with Staggered Adoption. Working paper, Universidad Carlos III de Madrid. http://dx.doi.org/10.2139/ssrn.3484635

St. Clair, T., Cook, T. D., Hallberg, K. (2014). Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison With a Randomized Experiment. *American Journal of Evaluation*, 35(3), 311-327. https://doi.org/10.1177/1098214014527337

St. Clair, T., Hallberg, K., & Cook, T. (2016). The Validity and Precision of the Comparative Interrupted Time-Series Design: Three Within-Study Comparisons. *Journal of Educational and Behavioral Statistics*, 41(3), 269-299. https://doi.org/10.3102/1076998616636854

Steiner, P. M. & Wong, V. C. (2018). Assessing Correspondence Between Experimental and Nonexperimental Estimates in Within-Study Comparisons. *Evaluation Review*, 42(2), 214-247. https://doi.org/10.1177/0193841X18773807

Unlu, F., Lauen, D. L. Fuller, S. C. Berglund, T., & Estrera, E. (2021). Can Quasi-Experimental Evaluations That Rely on State Longitudinal Data Systems Replicate Experimental Results? *Journal of Policy Analysis and Management*, 40(2) 572-613. https://doi.org/10.1002/pam.22295

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3-28. http://dx.doi.org/10.1257/jep.28.2.3

Visual Crossing Corporation. (2020). How Historical Weather Data Records are Created from Local Weather Station Observations. Visual Crossing. https://www.visualcrossing.com/resources/documentation/weather-data/how-historical-weather-data-records-are-created-from-local-weather-station-observations/

Visual Crossing Corporation. (2022). Visual Crossing Weather (2013-2016). [Data service]. Retrieved from https://www.visualcrossing.com/

Walker, D., Creaco, E., Vamvakeridou-Lyroudia, L., Farmani, R., Kapelan, Z. & Savić, D. (2015). Forecasting Domestic Water Consumption from Smart Meter Readings using Statistical Methods and Artificial Neural Networks. *Procedia Engineering*, 119, 1419-1428. https://doi.org/10.1016/j.proeng.2015.08.1002

Weller, D. L., Love, T. M., & Wiedmann, M. (2021). Interpretability Versus Accuracy: A Comparison of Machine Learning Models Built Using Different Algorithms, Performance Measures, and Features to Predict E. coli Levels in Agricultural Water. *Frontiers in Artificial Intelligence,* 4. https://doi.org/10.3389/frai.2021.628441

Wong, V. C. & Steiner, P. M. (2018). Designs of Empirical Evaluations of Nonexperimental Methods in Field Settings. *Evaluation Review*, 42(2), 176-213. https://doi.org/10.1177/0193841X18778918

Wong, V. C., Steiner, P. M., & Anglin, K. A. (2018). What Can Be Learned From Empirical Evaluations of Nonexperimental Methods? *Evaluation Review*, 42(2), 147-175. https://doi.org/10.1177/0193841X18776870

Zhang, L., Tang, H., & Bian, L. (2022). A Counterfactual Framework Based on the Machine Learning Method and Its Application to Measure the Impact of COVID-19 Local Outbreaks on the Chinese Aviation Market. *Aerospace*, 9(5), 250.

9. Appendix

Table A.1. Control Group Fixed Effects Regression Results

	Water use (cubic meters)
post_treat	0.5192
	(0.4156)
train	-0.1996
	(0.2637)
constant	0.6032
	(0.4773)
number of observations	12615
number of households	440

Table displays regression coefficients for estimates of monthly household water use, measured in cubic meters. Standard errors are bootstrapped according to the procedure described in Section 4.2.

*p<0.1, **p<.05, ***p<0.01

Table A.2. Variables Used in OLS Regression for Non-ML Approach

Category	Variable	Format
Household characteristics	$community$	dummies
	$community \cdot interviewer\ team$	interactions
	$household\ size$	continuous
	$years\ in\ home$	continuous
Education	$secondary\ school$	binary
	$CBWMO\ assembly\ participation$	binary
Income indicators	$income$	dummies
	$laptop$	binary
	$internet$	binary
	$owns\ home$	binary
Water devices	$showers$	continuous
	$baths$	continuous
	$kitchen\ faucets$	continuous
Temporal	$month$	dummies
Weather	$mean\ monthly\ temp$	continuous
	$rainy\ days$	continuous
	$uv\ index$	continuous

Equation A.1. Non-ML Fixed Effects Prediction Error Regression

$$Y_{it} - \hat{Y}_{it} = \beta_0 + \beta_1 \cdot post\_treat_{it} + household_i + month_j + \mu_{it}$$

Equation A.2. Single Interrupted Time Series Model

$$Y_{it} = \beta_0 + \beta_1 \cdot time_t + \beta_2 \cdot post\_treat_{it} + \beta_3 \cdot time\_since_{it} + \mathbf{X}\boldsymbol{\beta} + \epsilon_{it}$$
$$\epsilon_{it} = \rho \cdot \epsilon_{it-1} + \mu_{it}$$

$time_t$: months since beginning of period (begins at 1)
$post\_treat_{it}$: post-treatment dummy
$time\_since_{it}$: months since treatment (begins at 0)