Estimating Treatment Effects of a Residential Demand Response Program Using Non-Experimental Data

Datong P. Zhou
University of California, Berkeley
datong.zhou@berkeley.edu

Maximilian Balandat
University of California, Berkeley
balandat@berkeley.edu

Claire J. Tomlin
University of California, Berkeley
tomlin@berkeley.edu

Abstract*—Residential Demand Response has emerged as an instrument of the modern smart grid to alleviate supply and demand imbalances of electricity. Utilizing their flexibility of electricity demand, residential households are offered monetary incentives to temporarily reduce energy consumption during times when the grid is strained due to a supply shortage. In this paper, we estimate the magnitude of reductions of 1,025 residential households in California, serviced by the three main utilities PG&E, SDG&E, and SCE, using time-series regression on hourly smart meter data. By predicting the counterfactual consumption during Demand Response periods, which is the hypothetical consumption had there been no intervention, we find a user-averaged reduction of 0.12 kWh (8.6% of the mean consumption) per hour. The reduction is heterogeneous on a temporal and geographic level, as we find the largest reductions to occur in the early afternoon and early evening, as well as a positive correlation with ambient air temperature. Further, our findings suggest that households with automated smart-home devices, which can be automatically shut off during Demand Response periods, show a larger percentage reduction of the mean (12%) compared to households without (8.3%).

I. Introduction

Following the 1970s energy crisis, programs for demand-side management (DSM) were introduced on a global scale. Such programs seek to temporarily reduce consumers' electricity demand through financial incentive schemes during periods of electricity supply shortage. These programs are enabled by the integration of information and communications technology in the electric grid, characterizing the smart grid.

Energy supply is highly inelastic due to the slowness of power plants' output adjustment, which causes small increases and decreases in demand to result in a price boom or bust, respectively. This issue is exacerbated by the variable nature of electricity demand (mainly influenced by ambient temperatures [1]), prohibitively costly energy storage, and steady growth of renewable - and volatile - electricity generation. Despite the fact that electric utilities and generating companies hedge against such price fluctuations through long-term contracts, a large portion of electricity remains to be procured through the wholesale electricity market. Since utilities are obligated to serve end-users with electricity at a quasi-fixed tariff at all times, e.g. Time-of-Use pricing, they have to bear price risks in a volatile market. Therefore, DSM to reduce demand during peak periods is also an attempt to protect utilities against such price risks by partially relaying them to end-users, which increases market efficiency according to the economic consensus [2].

DSM describes a set of interventions aiming to affect customer behavior on different scales of application and time [3]. In this paper, we focus on individual households and short-term behavioral interventions during hours of peak demand or shortages of electricity supply, when demand reductions can counteract high electricity prices reflected by Locational Marginal Prices (LMPs) [4].

The California Public Utilities Commission (CPUC) has launched a Demand Response Auction Mechanism (DRAM) in July 2015 [5] which requires utilities to procure a minimum monthly amount of reduction capacity from Demand Response (DR) aggregators. The real-time market determines electricity prices by matching demand and utilities' supply curves subject to the procured capacity. A utility whose bid is cleared then asks the DR provider to incentivize its customers to temporarily reduce their consumption relative to their projected consumption without intervention. This is the counterfactual, referred to in this context as *baseline*, based on which compensations for (non-)fulfilled reductions are settled: If the consumer uses less (more) energy than the baseline, she receives a reward (incurs a penalty). In a similar fashion, if the aggregator falls short of delivering the promised load reduction, it incurs a penalty. For a profit-maximizing bid, the DR provider needs to estimate the counterfactual consumption as precisely as possible, among other aspects such as the LMP and the elasticity of users' demand in response to incentives.

We remark that the estimation of the actually delivered reduction, both on the household and aggregation level, is arguably the most critical component of the DR bidding process. If the reductions are estimated with a biased counterfactual, either the DR provider or the utility clearing the bids is systematically discriminated against. If the baseline is unbiased but plagued by high variance, the profit settlement is highly volatile. Existing baselines employed by major power grid operators in the United States (e.g. California Independent System Operator (CAISO), New York ISO) are calculated with simple arithmetic averages of previous observations [6] and therefore are inaccurate. Improving on such baselines by using more accurate estimators, while maintaining unbiasedness, is a central contribution of this paper.

A. Contributions

In this paper, we estimate the causal effect of a residential DR program in California on the reduction of electricity consumption of 1025 individual households serviced by the three main electric utilities in California (Pacific Gas & Electric (PG&E), San Diego Gas & Electric (SDG&E), Southern California Edison (SC&E)). The observational data is provided by the company `OhmConnect, Inc.` [7], headquartered in the San Francisco Bay Area, and is being held under a confidentiality agreement. We find an average reduction of 8.6% of the mean consumption and a 78.6% response rate to DR (users who reduced consumption) using the Hodges-Lehmann-Estimator. Further, we discover notable geographic and temporal heterogeneity among users. That is, the largest estimated reductions occur during early afternoon hours and early evenings, as well as in regions with warmer climate, suggesting that air conditioning units play a decisive role in DR programs. Lastly, households equipped with automated, smart-home devices that are remotely shut off during DR periods show an interesting anomaly: Despite reducing less energy (0.118 kWh) than households without smart devices (0.123 kWh), their relative reduction is higher (12% vs. 8.3% of mean consumption, respectively), indicating automated households, on average, consume less energy than non-automated ones.

B. Related Work

The rapid growth of collected user data has spurred research at the intersection of machine learning and economics with the goal of estimating treatment effects of an intervention in situations where Randomized Controlled Trials (RCTs), the experimental standard, are infeasible to conduct, e.g. due to budget or ethical constraints. The general idea is to partition observations under treatment and control in order to fit a nominal model on the latter set, which, when applied on the treatment set, yields counterfactual estimates (baselines), from which the treatment effect is computed by subtracting out actual observed treatment outcomes. Examples for such nominal models are found in [8], who evaluates welfare effects of home automation by calculating the Kolmogorov-Smirnov Statistics between users, which are then used as weights for kernel-based non-parametric regression. In [9], a convex combination of US states is computed as the counterfactual estimate for tobacco consumption to estimate the effect of a tobacco control program in California on tobacco consumption. In [10], [11], the estimators are random forests trained by recursive partitioning of the feature space and novel cross-validation criteria. [12] develops Bayesian structural time series models combined with a Monte-Carlo sampling method for treatment effect inference of market interventions.

Fitting an estimator on smart meter time-series is essentially a short-term load forecasting (STLF) problem, whose goal is to fit estimators on observed data to predict future consumption with the highest possible accuracy. Within STLF, tools employed are ARIMA models with a seasonal component [13] and classic regression models where support vector regression (SVR) [14] and neural networks [15] yield the highest accuracy. A comprehensive comparison between ML techniques for forecasting and differing levels of load aggregation can be found in [16].

In the context of data mining smart meter data, much of the existing work focuses on disaggregation of energy consumption to identify contributions of discrete appliances from the total observed consumption [17], [18], and to learn consumption patterns [19], [20], [21] using clustering approaches [22]. Studies in applied economics typically emphasize the estimation of *average* treatment effects of experimental interventions. To increase precision of the estimates, the employed regression models often employ unit-level fixed effects [23] as an implicit way of training models for the consumption of individual consumers. In this work, we make these user-level models explicit, allowing for more general ML techniques. Importantly, our approach is original as it permits to perform causal inference on the level of *individual* treatment effects by employing estimators from STLF. To the best of our knowledge, this paper is the first of its kind to analyze the potential of Demand Response interventions on a residential level, combining ideas at the intersection of causal inference from econometrics and Machine Learning for estimation.

II. Demand Response Mechanism

According to DRAM [5], electric utilities are obligated to offer “demand flexibility” through Demand Response Providers (DRPs). Utilities solicit bids from DRPs and accept the highest ones up to a monthly target capacity. In the real-time wholesale electricity market, the utility submits supply bids including these acquired capacities, which, when cleared, have to be delivered by the DRP under contract over a contractually specified period of time. The DRP does so by eliciting reductions among a suitable chosen subset of its residential end-use customers by offering them a monetary incentive. Such an aggregation of users is also known as a *Proxy Demand Resource* (PDR) [6] product. The DRP receives a payment from the wholesale market for each unit of reduction up to its original capacity bid, but incurs a shortfall penalty for each unit of unfulfilled obligation. Figure 1 illustrates the interaction between all agents.

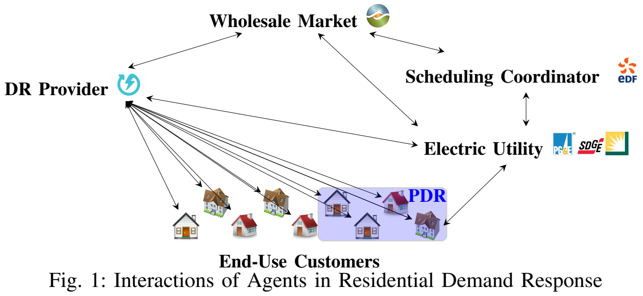

Fig. 1: Interactions of Agents in Residential Demand Response

We focus on the DRP-User interaction and in particular answer the question of how to measure and quantify reductions of end-users' electricity consumption in response to monetary incentives. Our data set consists of DR events of length one hour. Specifically, users receive notifications of a DR event up to 20 minutes into an hour, which lasts until the end of the hour. Further, users were compensated proportional to their reductions, which, however, was revealed to them only after the end of the DR event. The effect we are hence analyzing is the impact of notifying users of a DR event on the reduction of electricity consumption.

III. Treatment Effect Estimation

A. Potential Outcomes Framework

To estimate the effect of the DR intervention program, we adopt the *potential outcomes* framework introduced by Rubin (1974) [24]. Let $\mathcal{I} = \{1, \ldots, n\}$ denote the set of users. The indicator $D_{it} \in \{0, 1\}$ encodes the fact whether or not user i received DR treatment at time t. Each user is endowed with a consumption time series $\mathbf{y}_i = \{y_{i1}, \ldots, y_{i\tau}\}$ and associated covariates $X_i = \{\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i\tau}\} \in \times_{i=1}^{\tau} \mathcal{X}_i$, $\mathcal{X}_i \subset \mathbb{R}^{n_x}$, where time is indexed by $t \in \mathbb{T} = \{1, \ldots, \tau\}$ and n_x is the dimension of the covariate space $\mathcal{X}_i$. Let y_{it}^0 and y_{it}^1 denote user i's electricity consumption at time t for $D_{it} = 0$ and $D_{it} = 1$, respectively. Let $\mathcal{C}_i$ and $\mathcal{T}_i$ denote the set of control and treatment times for user i. That is,

$$\mathcal{C}_i = \{t \in \mathbb{T} \mid D_{it} = 0\}, \quad \mathcal{T}_i = \{t \in \mathbb{T} \mid D_{it} = 1\}. \tag{1}$$

The number of treatment hours is much smaller than the number of non-treatment hours. Thus $0 < |\mathcal{T}_i|/|\mathcal{C}_i| \ll 1$.

Further, let $\mathcal{D}_{i,t}$ and $\mathcal{D}_{i,c}$ denote user i's covariate-outcome pairs of treatment and control times, respectively. That is,

$$\mathcal{D}_{i,t} = \{(\mathbf{x}_{it}, y_{it}) \mid t \in \mathcal{T}_i\}, \quad \mathcal{D}_{i,c} = \{(\mathbf{x}_{it}, y_{it}) \mid t \in \mathcal{C}_i\}. \tag{2}$$

The one-sample estimate of the treatment effect on user i at time t, given the covariates $\mathbf{x}_{it} \in \mathbb{R}^{n_x}$, is

$$\beta_{it}(\mathbf{x}_{it}) := y_{it}^1(\mathbf{x}_{it}) - y_{it}^0(\mathbf{x}_{it}) \quad \forall\ i \in \mathcal{I},\ t \in \mathbb{T}, \tag{3}$$

which varies across time, the covariate space, and the user population. Marginalizing this one-sample estimate over the set of treatment times $\mathcal{T}_i$ and the covariate space $\mathcal{X}_i$ yields the user-specific Individual Treatment Effect (ITE) β_i

$$\beta_i := \mathbb{E}_{\mathcal{X}_i} \mathbb{E}_{t \in \mathcal{T}_i} \left[y_{it}^1 - y_{it}^0 \;\middle|\; \mathbf{x}_{it} \right] = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} y_{it}^1 - y_{it}^0. \tag{4}$$

The average treatment effect on the treated (ATT) follows from (4):

$$\text{ATT} := \mathbb{E}_{i \in \mathcal{I}}[\beta_i] = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} (y_{it}^1 - y_{it}^0). \tag{5}$$

As we cannot rule out selection bias among users who have chosen to subscribe to the DR program, (5) does not coincide with the average treatment effect (ATE) [25].

Lastly, the conditional average treatment effect on the treated (CATT) on $\tilde{\mathbf{x}}$ is obtained by marginalizing the conditional distribution of one-sample estimates (3) on $\tilde{\mathbf{x}}$ over all users and treatment times, where $\tilde{\mathbf{x}} \in \mathbb{R}^{\tilde{n}_x}$ is a subvector of $\mathbf{x} \in \mathbb{R}^{n_x}$, $0 < \tilde{n}_x < n_x$:

$$\text{CATT}(\tilde{\mathbf{x}}) := \mathbb{E}_{i \in \mathcal{I}} \mathbb{E}_{t \in \mathcal{T}_i} \left[(y_{it}^1 - y_{it}^0) \;\middle|\; \tilde{\mathbf{x}}_{it} = \tilde{\mathbf{x}} \right]. \tag{6}$$

The CATT captures heterogeneity among users, e.g. with respect to specific hours of the day, the geographic distribution of users, the extent to which a user possesses "smart home" appliances, group or peer effects, etc. To rule out the existence of unobserved factors that could influence the assignment mechanism generating the complete observed data set $\{(\mathbf{x}_{it}, y_{it}, D_{it}) \mid i \in \mathcal{I}, t \in \mathbb{T}\}$, we make the following standard assumptions:

Assumption 1 (Unconfoundedness of Treatment Assignment)**.** *Given the covariates $\{\mathbf{x}_{it}\}_{t \in \mathbb{T}}$, the potential outcomes are independent of treatment assignment:*

$$(y_{it}^0, y_{it}^1) \perp D_{it} \mid \mathbf{x}_{it} \quad \forall i \in \mathcal{I}, t \in \mathbb{T}. \tag{7}$$

Assumption 2 (Stationarity of Potential Outcomes)**.** *Given the covariates $\{\mathbf{x}_{it}\}_{t \in \mathbb{T}}$, the potential outcomes are independent of time, that is,*

$$(y_{it}^0, y_{it}^1) \perp t \mid \mathbf{x}_{it} \quad \forall i \in \mathcal{I}, t \in \mathbb{T}. \tag{8}$$

Assumption 1 is the "ignorable treatment assignment" assumption introduced by Rosenbaum and Rubin [26]. Under this assumption, the assignment of DR treatment to users is implemented in a *randomized* fashion, which allows the calculation of unbiased ATTs (5) and CATTs (6). Assumption 2, motivated by the time-series nature of the observational data, ensures that the set of observable covariates $\{\mathbf{x}_{it} \mid t \in \mathbb{T}\}$ can capture seasonality effects in the estimation of the potential outcomes. That is, the conditional distribution of the potential outcomes, given covariates, remains constant.

The *fundamental problem of causal inference* [27] refers to the fact that either the treatment or the control outcome can be observed, but never both (granted there are no missing observations). That is,

$$y_{it} = y_{it}^0 + D_{it} \cdot (y_{it}^1 - y_{it}^0) \quad \forall\ t \in \mathbb{T}. \tag{9}$$

Thus, the ITE (4) is not identified because one and only one of both potential outcomes is observed, namely $\{y_{it}^1 \mid t \in \mathcal{T}_i\}$ for the treatment times and $\{y_{it}^0 \mid t \in \mathcal{C}_i\}$ for the control times. It therefore becomes necessary to estimate counterfactuals. Consider the following model for the estimation of such counterfactuals:

$$y_{it} = f_i(\mathbf{x}_{it}) + D_{it} \cdot \beta_{it}(\mathbf{x}_{it}) + \varepsilon_{it}, \tag{10}$$

where ε_{it} denotes noise uncorrelated with covariates and treatment assignment. The conditional mean function $f_i(\cdot) : \mathbb{R}^{n_x} \mapsto \mathbb{R}$ pertains to $D_{it} = 0$. To obtain an estimate for $f_i(\cdot)$, denoted with $\hat{f}_i(\cdot)$, control outcomes $\{y_{it}^0 \mid t \in \mathcal{C}_i\}$ are first regressed on their observable covariates $\{\mathbf{x}_{it} \mid t \in \mathcal{C}_i\}$. In a second step, the counterfactual $\hat{y}_{it}^0$ for any $t \in \mathcal{T}_i$ can be estimated by evaluating $\hat{f}_i(\cdot)$ on its respective covariate vector $\mathbf{x}_{it}$. Finally, subtracting $\hat{y}_{it}^0$ from y_{it}^1 isolates the one-sample estimate $\beta_{it}(\mathbf{x}_{it})$, from which the user-specific ITE (4) can be estimated. Figure 2 illustrates this process of estimating the reduction during a DR event by subtracting the actual consumption y_{it}^1 from the predicted counterfactual $\hat{y}_{it}^0 = \hat{f}_i(\mathbf{x}_{it})$. Despite the fact that consumption can be predicted for horizons longer than a single hour, we restrict our estimators $f_i(\cdot)$ to a single hour prediction horizon.

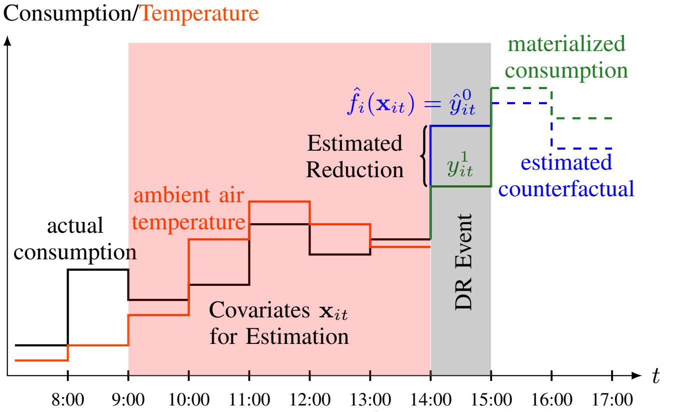

Fig. 2: Estimation of the Counterfactual $\hat{y}^0_{it}$ using Treatment Covariates $\mathbf{x}_{it}$ and Predicted Reduction $\hat{y}^0_{it} - y^1_{it}$

To estimate the conditional mean function $f_i(\cdot)$, we use the following, classical regression methods [28], referred to as *estimators*:

E1: Ordinary Least Squares Regression (OLS)

E2: L1 Regularized (LASSO) Linear Regression (L1)

E3: L2 Regularized (Ridge) Linear Regression (L2)

E4: k-Nearest Neighbors Regression (KNN)

E5: Decision Tree Regression (DT)

E6: Random Forest Regression (RF)

DT (E5) and RF (E6) follow the procedure of Classification and Regression Trees [29]. We compare estimators (E1)-(E6) to the CAISO 10-in-10 Baseline (BL) [6], which, for any given hour on a weekday, is calculated as the mean of the hourly consumptions on the 10 most recent business days during the selected hour. For weekend days and holidays, the mean of the 4 most recent observations is calculated. This BL is further adjusted with a *Load Point Adjustment*, which corrects the BL by a factor proportional to the consumption three hours prior to a DR event, excluding the hour immediately prior to the event.

B. Nonparametric Signed Rank Test

Rather than naively computing the differences in means between the treatment observations $\{y^1_{it} \mid t \in \mathcal{T}_i\}$ and their estimated counterfactuals $\{\hat{y}^0_{it} \mid t \in \mathcal{T}_i\}$, a nonparametric comparison between these sets admits p-values and coverage probabilities for confidence intervals without requiring assumptions on the underlying data generating process, while being robust against outliers. The paired replicate nature of these sets as well as their relatively small size (which precludes the use of the Central Limit Theorem) calls for an analysis using signed ranks [30]. We estimate ITEs with the Hodges-Lehmann Estimator (HLM) that is associated with the Wilcoxon Signed Rank Statistic. The Hodges-Lehmann Estimator is intrinsically related to the Wilcoxon Signed Rank Test with null hypothesis and its corresponding alternative

$$H_0 : \beta_i = 0, \tag{11a}$$
$$H_1 : \beta_i \neq 0, \tag{11b}$$

which constitutes a two-sided test at significance level α.

The confidence interval $[\underline{\theta}_i, \overline{\theta}_i]$ admitted by the Wilcoxon Signed Rank Test corresponds to the range of $\tilde{\beta}_i$ for which it does not reject the modified null hypothesis $H_0 : \beta = \tilde{\beta}_i$ with corresponding alternative $H_1 : \beta \neq \tilde{\beta}_i$ at significance level α. Thus, (11a) is rejected at confidence level $1 - \alpha$ if

- $\underline{\beta}_i \leq \overline{\beta}_i < 0$: (11a) is rejected at the lower tail (that is, user i *reduces* consumption significantly), or
- $0 < \underline{\beta}_i \leq \overline{\beta}_i$: (11a) is rejected at the upper tail (that is, user i *increases* consumption significantly).

IV. Data and Data Preparation

A. Aggregate Statistics

Our data set consists of $|\mathcal{I}| = 5000$ residential customers in California serviced by the three largest investor owned utilities (PG&E, SCE, and SDG&E). For each user, we have its ZIP code, hourly smart meter time series data of varying lengths, and timestamps of hourly DR events. Figure 3 shows the geographic distribution of users across California. To train estimators (E1)-(E6) on non-DR periods, sufficiently long historical consumption data is required, and so we drop all users with less than seven months of data. Further, users with net energy metering (NEM) tariffs (due to rooftop solar panels) are removed. After these operations, 1025 users remain, with a subset of 83 users that owns at least one "smart home" device which can be remotely shut off by the DR provider, given the users' approval to do so. Devices include thermostats (`Nest`, `Honeywell`, `Schneider Electric Wiser`, `Ecobee`) and smart plugs (`TP-Link`).

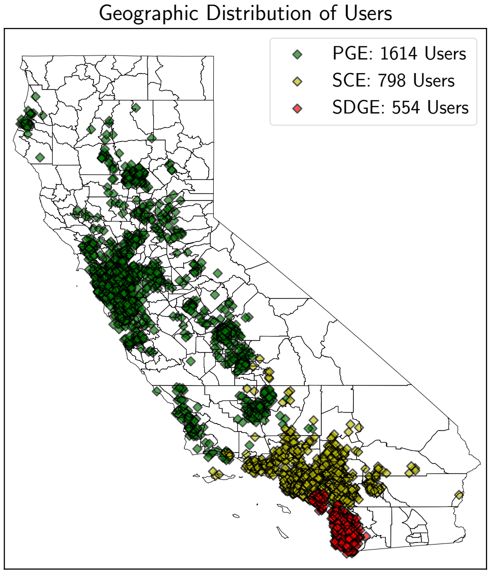

Fig. 3: Geographic Distribution of Users

B. Data Preprocessing

Hourly measurements of ambient air temperature are scraped from the publicly accessible California Irrigation Management Information System (CIMIS) [31]. As there are fewer weather stations than distinct user ZIP codes, we linearly interpolate user-specific temperatures at their ZIP codes from the two closest weather stations in latitude and longitude by calculating geodesic distances with Vincenty's formulae [32].

Moreover, since users tend to exhibit a temporary increase in consumption in the hours following the DR intervention [3], we remove $n_r = 10$ hourly observations following each DR event in order to prevent estimators (E1)-(E6) from learning from such spillover effects.

For each user, the cleaned hourly consumption time series is then split into the following two sets:

1) $\mathcal{D}_{i,t}$, the set of treatment data;
2) $\mathcal{D}_{i,c} = \mathcal{D}_{i,tr} \cup \mathcal{D}_{i,pl} \cup \mathcal{D}_{i,syn}$, the set of control data;
   - $\mathcal{D}_{i,tr} \subset \mathcal{D}_{i,c}$, the training data set;
   - $\mathcal{D}_{i,pl} \subset \mathcal{D}_{i,c}$, the placebo treatment data set;
   - $\mathcal{D}_{i,syn} \subset \mathcal{D}_{i,c}$, the synthetic treatment data set,

where 1% each of $\mathcal{D}_{i,c}$ are randomly allocated to $\mathcal{D}_{i,pl}$ and $\mathcal{D}_{i,syn}$ according to the empirical distribution of DR events, which is depicted in Figure 4. Let $\mathcal{S}_i$ and $\mathcal{P}_i$ denote the ensuing sets of user i's synthetic treatment and placebo treatment times, respectively. The remaining 98% of $\mathcal{D}_{i,c}$ are allocated to the training set $\mathcal{D}_{i,tr}$, on which outcome estimators (E1)-(E6) are fitted.

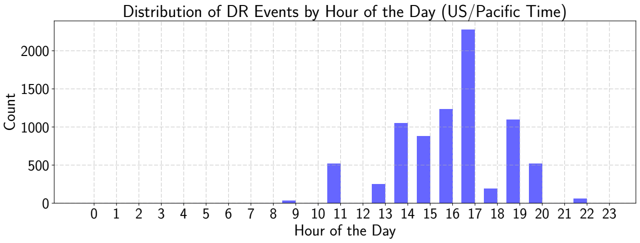

Fig. 4: Distribution of DR Events Across Hours of the Day

1) Placebo Treatments: The placebo treatment set $\mathcal{D}_{i,pl}$ is used to test the accuracy and unbiasedness of estimators (E1)-(E6) by creating treatment outcomes $\{y^1_{it} \mid t \in \mathcal{P}_i\}$ with a zero treatment effect (hence the name "placebo treatment"). These outcomes are to be recovered by the counterfactual estimates $\{\hat{y}^0_{it} \mid t \in \mathcal{P}_i\}$. The one-sample estimation errors

$$\{\hat{y}^0_{it} - y^1_{it} \mid t \in \mathcal{P}_i\} \tag{12}$$

of an unbiased estimator should be centered around zero. The more accurate the estimator, the lower we expect the sample variance of (12) to become.

2) Synthetic Treatments: The synthetic treatment set $\mathcal{D}_{i,syn}$ is used as a set of ground truth counterfactuals $\{y^0_{it} \mid t \in \mathcal{S}_i\}$ for which treatment outcomes $\{y^1_{it} \mid t \in \mathcal{S}_i\}$ are synthetically generated. We assume a constant ITE (4), $-1 \leq \beta_1 = \ldots = \beta_N =: \beta \leq 0$, across all synthetic times and the covariate space for each user i, as a percentage of user i's mean counterfactual consumption. The one-sample reductions (3) are varied around the mean reduction through Gaussian noise with an appropriate standard deviation:

$$\mu_i := \frac{1}{|\mathcal{S}_i|}\sum_{t\in\mathcal{S}_i} y^0_{it}, \tag{13a}$$

$$y^1_{it} \leftarrow y^0_{it} + \beta_{it}, \quad \beta_{it} \sim \mathcal{N}(\beta\mu_i, \sigma^2) \quad \forall\, t \in \mathcal{S}_i. \tag{13b}$$

Since β_{it} is random in σ^2, the realized ITE β_i is distributed according to $\beta_i \sim \mathcal{N}(\beta, \sigma^2/(\mu_i^2|\mathcal{S}_i|))$, which follows from (13b) and noting that $\{y_{it}\}_{t\in\mathcal{S}_i}$ are independent random variables. Using this semi-synthetic treatment data, one can evaluate the ability of the estimators to recover the generated ITE $\beta\mu_i$ (non-normalized) and β (normalized). The sample variance of the ITE estimation errors will again serve as a measure for the predictive power of an estimator, similar to the estimation error of placebo treatments.

3) Training of Estimators: The training data $\mathcal{D}_{i,tr}$ is used to estimate the conditional mean function $\hat{f}_i(\cdot)$ (10) of models (E1)-(E6) with standard k-fold cross-validation on the following covariates:

- Hour of the day, day of the week, and month of the year as categorical variables,
- Previous $n_{ar} = 5$ hourly measurements of ambient air temperature, and
- $\mathbf{y}_{i,t-n_{ar}:t-1} := \{y_{i,t-n_{ar}}, \ldots, y_{i,t-1}\}$, i.e. the previous n_{ar} hourly consumption values.

Observations for which the autoregressive term $\mathbf{y}_{i,t-n_{ar}:t-1}$ does not exist are dropped from the data set. More specifically, for any two treatment observations y_{it} and $y_{i\tilde{t}}, \tilde{t} > t$, we must have that $\tilde{t} - t > n_{ar} + n_r$, i.e. we only include treatment observations which are separated by at least $n_{ar} + n_r$ non-treatment hours. For the choice $n_{ar} = 5$ and $n_r = 10$, which we stick to throughout this paper, this requirement is fulfilled for more than 97% of all DR events. To reduce the propagation of model bias into the estimation of treatment effects, we empirically de-bias estimators by subtracting the empirical bias, which is the difference in means between the observed control outcomes and their predictions, from all estimated counterfactuals:

$$\hat{y}^0_{it} \leftarrow \hat{y}^0_{it} - \frac{1}{|\mathcal{C}_i|}\sum_{k\in\mathcal{C}_i}(\hat{y}^0_{ik} - y^0_{ik}) \quad \forall\, t \in \mathcal{T}_i. \tag{14}$$

Although (14) leads to an increase of variance of counterfactual estimates, the reasoning behind this operation is that an unbiased estimator provides a fair economic settlement for DR reductions. If the estimator were biased in favor of the consumer, then the user, in expectation, would receive an additional payment proportional to the bias each time a DR event is called despite not having actually reduced his consumption by the amount of bias. Likewise, an estimator biased in favor of the utility results in the opposite effect.

V. Simulation Results

A. Results on Control Data

1) One-Sample Predictions: Table I provides the sample bias and standard deviation of the distribution of one-sample prediction errors (12) on the placebo treatment set. It becomes clear that RF outperforms all other estimators as it is has the smallest sample standard deviation. The performance of L1, L2, and OLS are similar to each other, yet worse then RF, indicating that the training data is of sufficient size such that overfitting is not a concern. The performance of KNN lies between L1/L2/OLS and BL. The CAISO BL performs worst. As the estimators were calibrated with the de-biasing

Sample Mean and Variance on Placebo Treatment Set $\mathcal{D}_{i,pl}$			
Method	Bias	St. Dev.	median MAPE [%]
RF	0.00280	**0.34460**	**30.779**
OLS	0.00157	0.35981	35.088
L1	0.00184	0.35969	34.945
L2	0.00153	0.35977	35.079
DT	−8.26e-05	0.40386	35.461
KNN	−0.00129	0.41011	41.341
BL	0.00684	0.49550	50.496

TABLE I: MAPE, Sample Bias and Standard Deviation of Placebo Predictions for Estimators (E1)-(E6) and BL

operation (14), the one-sample estimation errors (whose mean is the bias) for all estimators varies insignificantly around zero. Table I also provides the median of the set of Mean Absolute Percentage Errors (MAPE) across all users and for all estimators (E1)-(E6). The MAPE for a given user i is

$$\text{MAPE} = \frac{1}{|\mathcal{V}_i|}\sum_{t\in\mathcal{V}_i}\left|\frac{\hat{f}_i(\mathbf{x}_{it}) - y^0_{it}}{y^0_{it}}\right| \cdot 100\%, \qquad (15)$$

where $\mathcal{V}_i \subset \mathcal{C}_i$ is a subset of the set of training times used for validation of the estimators during the training step. Using standard k-fold cross validation on the training data set $\mathcal{D}_{i,tr}$ (i.e. we chose $k = 10$), $\mathcal{V}_i$ can be interpreted as the set of time indices in the holdout set of any given fold.

2) Estimation of ITEs: Figure 5 shows the distribution of estimated normalized ITEs $\{\hat{\beta}_i\}_{i\in\mathcal{I}}$ generated in (13a) and (13b) across all users, for selected estimators, and for two different ground truth ITEs $\beta_i \in \{-0.01, -0.15\}$. Each ITE draw is obtained from a randomly drawn subset $\mathcal{M}_i \subset \mathcal{S}_i$, where we chose $|\mathcal{M}_i| = 25$. As in Table I, RF outperforms estimators (E1)-(E5) and BL, as the histograms around the sample mean become wider as we move to the more inaccurate estimators towards the bottom of the figure.

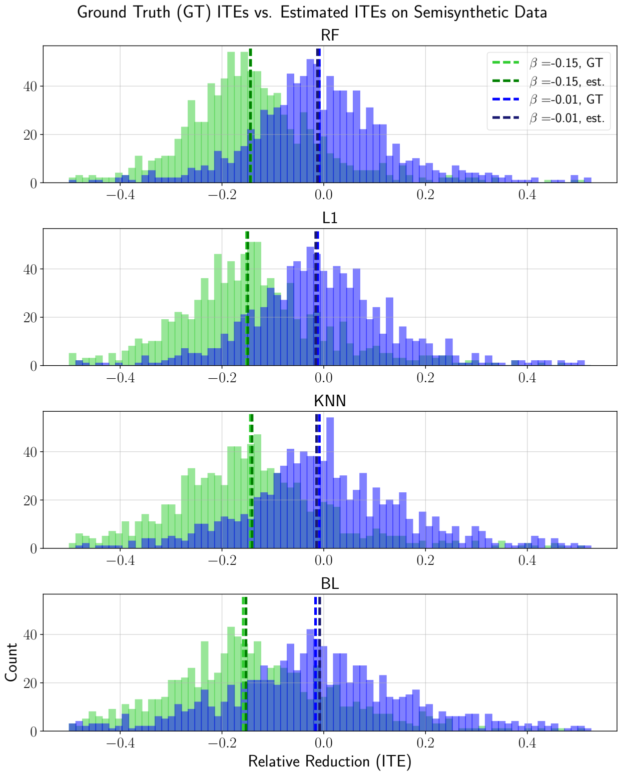

Fig. 5: Distribution of ITE Estimates Across Users and Estimators (E6), (E2), (E4), and CAISO BL. Green: $\beta_i = -0.15 \cdot \mu_i$, Blue: $\beta_i = -0.01 \cdot \mu_i$.

3) Economic Implications: From an economic perspective, the following misclassification errors are costly:

- If $y^0_{it} > y^1_{it}$, but $\hat{y}^0_{it} < y^1_{it}$ is predicted, user i receives no credit despite having actually reduced consumption.
- Conversely, if $y^0_{it} < y^1_{it}$, but $\hat{y}^0_{it} > y^1_{it}$ is predicted, user i is over-credited despite an increase in consumption.

To quantify these errors, we calculate the population-wide conditional means of estimated reductions (increases) on the synthetic treatment set, denoted with $\mu^{\downarrow\uparrow}$ and $\mu^{\uparrow\downarrow}$, respectively, given that the user actually increased (reduced):

$$\mu^{\downarrow\uparrow} = \frac{\sum_{i\in\mathcal{I}}\sum_{t\in\mathcal{S}_i}(\hat{y}^0_{it} - y^1_{it}) \cdot \mathbf{1}_{\hat{y}^0_{it}>y^1_{it}} \cdot \mathbf{1}_{y^1_{it}>y^0_{it}}}{\sum_{i\in\mathcal{I}}\sum_{t\in\mathcal{S}_i}\mathbf{1}_{\hat{y}^0_{it}>y^1_{it}} \cdot \mathbf{1}_{y^1_{it}>y^0_{it}}}, \qquad (16a)$$

$$\mu^{\uparrow\downarrow} = \frac{\sum_{i\in\mathcal{I}}\sum_{t\in\mathcal{S}_i}(y^1_{it} - \hat{y}^0_{it}) \cdot \mathbf{1}_{y^1_{it}>\hat{y}^0_{it}} \cdot \mathbf{1}_{y^0_{it}>y^1_{it}}}{\sum_{i\in\mathcal{I}}\sum_{t\in\mathcal{S}_i}\mathbf{1}_{y^1_{it}>\hat{y}^0_{it}} \cdot \mathbf{1}_{y^0_{it}>y^1_{it}}}. \qquad (16b)$$

Let $\downarrow\uparrow$ and $\uparrow\downarrow$ denote the population-wide percentage of cases where a user is falsely given credit (not given credit) despite having increased (decreased) consumption:

$$\downarrow\uparrow = \frac{\sum_{i\in\mathcal{I}}\sum_{t\in\mathcal{S}_i}\mathbf{1}_{\hat{y}^0_{it}>y^1_{it}} \cdot \mathbf{1}_{y^1_{it}>y^0_{it}}}{\sum_{i\in\mathcal{I}}|\mathcal{S}_i|}, \qquad (17a)$$

$$\uparrow\downarrow = \frac{\sum_{i\in\mathcal{I}}\sum_{t\in\mathcal{S}_i}\mathbf{1}_{y^1_{it}>\hat{y}^0_{it}} \cdot \mathbf{1}_{y^0_{it}>y^1_{it}}}{\sum_{i\in\mathcal{I}}|\mathcal{S}_i|}. \qquad (17b)$$

We define the misclassification score R as the sum of conditional means of falsely estimated reductions (increases) (16a),(16b) multiplied by the percentage these cases occur, i.e. $R = \downarrow\uparrow \cdot \mu^{\downarrow\uparrow} + \uparrow\downarrow \cdot \mu^{\uparrow\downarrow}$. R is interpreted as the amount of falsely allocated credit per DR event and user due to estimation errors. Table II reports these metrics for a ground truth ATT of $\beta = 0.01$. As expected, RF achieves the lowest misclassification score, namely about 38% less than BL.

Misclassifications of Estimators for ATT $\beta = 0.01$						
	$\downarrow\uparrow$	$\uparrow\downarrow$	$\downarrow\uparrow + \uparrow\downarrow$	$\mu^{\downarrow\uparrow}$	$\mu^{\uparrow\downarrow}$	R
RF	.109	**.213**	**.322**	.153	**.408**	**.1036**
OLS	.122	.215	.337	.153	.435	.1122
L1	.123	.222	.345	**.143**	.444	.1162
L2	.122	.215	.337	.153	.435	.1122
DT	**.098**	.244	.342	.175	.416	.1187
KNN	.126	.247	.373	.209	.462	.1404
BL	.119	.236	.355	.241	.584	.1665

TABLE II: Misclassification Table for (E1)-(E6) and CAISO BL

B. Analysis of Observational Data

Since the Random Forest estimator (E6) has shown the highest accuracy among estimators (E1)-(E6) and in particular the CAISO BL, we restrict our analysis to this estimator in the remainder of the paper. Figure 7 shows Hodges-Lehmann estimates for the ITEs of those 515 users with at least 10 DR events together with their confidence intervals based on RF predictions. A subset of 43 users has at least one automated smart home device, for which the confidence intervals in Figure 7 are drawn in yellow.

At confidence level $1 - \alpha = 0.9$, 78.6% of all users are estimated to reduce their consumption, with 32.8% of the total

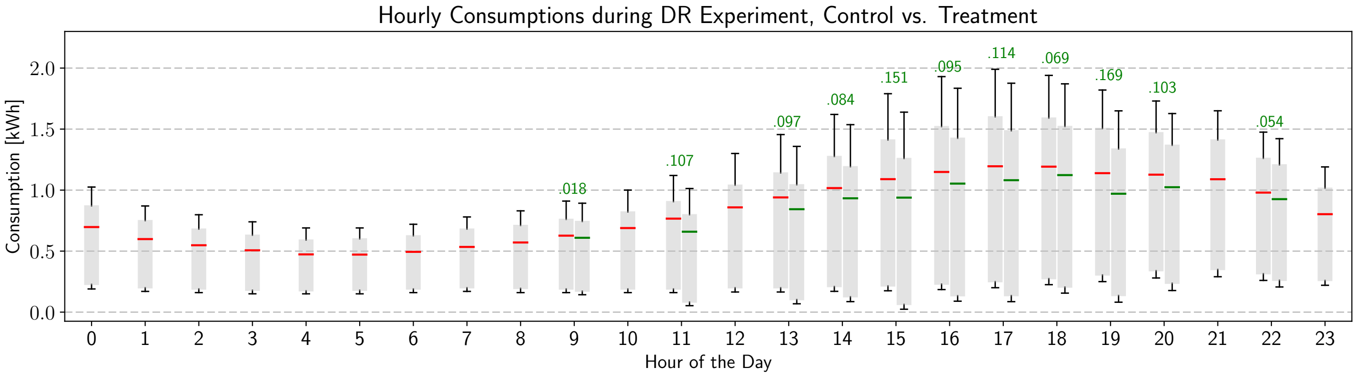

Fig. 6: CATT (6) by Hours of the Day on 515 Users with $\geq$ 10 DR Events. Red: Control Observations, Green: DR Estimates

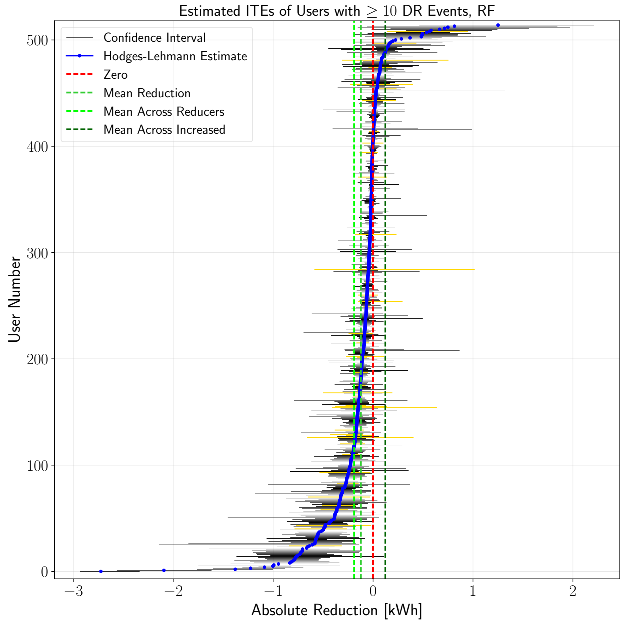

Fig. 7: ITEs and Confidence Intervals on 515 users with $\geq$ 10 DR Events, computed with Hodges-Lehmann Estimator and Wilcoxon Signed Rank Test, $1 - \alpha = 0.9$ Confidence Level

population reducing significantly. The remaining users have increased their consumption, with 1.5% of the total population having done so significantly. These are likely random artifacts - note that we expect 5% under the null hypothesis (11a) and at 90% confidence level.

1) Influence of Automation: Let A_i denote user i's number of automated "smart home" devices. Conditioning the ATT on the indicator whether or not a user has at least one automated device yields the CATT by automation status. Table III provides these CATT estimates. While the absolute reduction among $A \geq 1$ and $A = 0$ is similar in magnitude, their percentage reductions differ, indicating that automated users are more energy efficient to begin with. There is no notable difference in the estimated percentage of reducers.

ATT and CATTs by Automation Status

	CATT/ATT		(C)ATT on reducers		
	kWh	% of mean	kWh	% of mean	% reducers
All	-0.123	-8.59	-0.190	-15.3	78.6
$A \geq 1$	-0.118	-12.0	-0.173	-22.1	81.4
$A = 0$	-0.123	-8.28	-0.189	-14.6	78.4

TABLE III: CATT Estimates on 43 Automated and 472 Non-Automated Users with at least 10 DR Events

2) Heterogeneity in Time: Figure 6 shows the range of observed consumptions between the first and last DR event across all users for control periods and the estimated CATTs by different hours of the day, where "17" denotes consumption between 4-5 pm. The CATTs for hours 9 and 22 are smallest, which agrees with expectations. The largest effects are found to occur in the late afternoon and early evening, potentially when most users are at home.

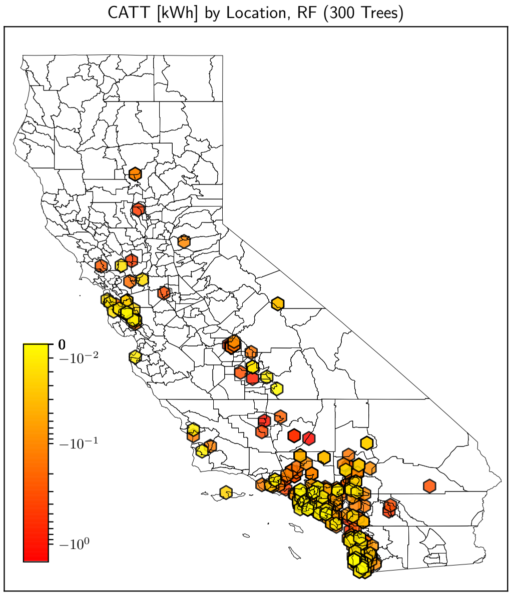

Fig. 8: CATT (6) by Geographic Region

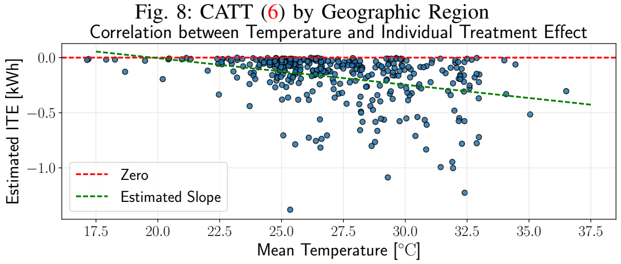

Fig. 9: Correlation between Ambient Air Temperature and Estimated ITE, estimated with RF (300 Trees)

3) Spatial Heterogeneity: Conditioning the ATT (5) on the ZIP codes of users yields the CATT by location. Figure 8 plots the location of each of the 515 users on a map of California, where the CATT is color coded. The highest CATTs are found in the inland areas of California (San Joaquin Valley), which are considerably warmer than the coastal areas. The positive correlation with ambient air temperature is supported also by Figure 9, which scatter plots the HLM estimates of ITEs against the average ambient air temperature during DR events. A simple t-Test on the paired ITE-temperature samples computes a p-value of 2.1e-11 for the null hypothesis H_0 : *ITE and temperature are uncorrelated*. This observation suggests a considerable impact of air conditioning units on the quantity of reduction during DR events.

VI. Conclusion

This paper estimates treatment effects of a residential Demand Response program on the reduction of energy consumption, using observational data provided by `OhmConnect, Inc` [7]. The Hodges-Lehmann Estimator in combination with the Wilcoxon Signed Rank Test is employed to construct non-parametric estimates and confidence intervals for reductions during DR events. Estimating the counterfactual consumption with classical Machine Learning regression methods, in particular Random Forest Regression, allows for a considerable improvement in predictive accuracy compared to the regulatory baselining standard set by the California Independent System Operator. By simulating users' responses on observed time series, it is shown that the Random Forest regressor reduces the extent of misclassifications in DR settlements compared to the CAISO baseline. An average treatment effect of -0.12 kWh per DR event and user is estimated across all treated users, which amounts to an 8.6% reduction from the mean consumption. 78.6% of the treatment population is estimated to respond to incentives by reducing consumption.

We further detect significant heterogeneity in time, automation status, and a positive correlation between the magnitude of reduction and ambient air temperature. To achieve external validity of the non-experimental treatment effect estimates proposed in this paper, we are currently conducting a Randomized Controlled Trial on more than 10,000 users in California. The DR events in this experiment include reward levels that are communicated to the users, which allows for an estimation of a *demand curve*. Lastly, we intend to investigate the welfare impact of Demand Response on end-users of electricity under a broader array of incentives, including non-monetary incentives and social comparisons.

References

[1] A. Pardo, V. Meneu, and E. Valor, "Temperature and Seasonality Influences on Spanish Electricity Load," *Energy Economics*, vol. 24, no. 1, pp. 55–70, 2002.

[2] S. Borenstein, "The Long-Run Efficiency of Real-Time Electricity Pricing," *The Energy Journal*, 2005.

[3] P. Palensky and D. Dietrich, "Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 381–388, 2011.

[4] Y.-Y. Hong and C.-Y. Hsiao, "Locational Marginal Price Forecasting in Deregulated Electricity Markets Using Artificial Intelligence," *IEE Proceedings - Generations, Transmission and Distribution*, vol. 149, no. 5, pp. 621–626, 2002.

[5] "Public Utilities Commission of the State of California: Resolution E-4728. Approval with Modifications to the Joint Utility Proposal for a Demand Response Auction Mechanism Pilot," July 2015.

[6] "California Independent System Operator Corporation (CAISO): Fifth Replacement FERC Electric Tariff," 2014.

[7] "OhmConnect, Inc. 2015," "https://www.ohmconnect.com", 2017.

[8] B. Bollinger and W. R. Hartmann, "Welfare Effects of Home Automation Technology with Dynamic Pricing," *Stanford University, Graduate School of Business Research Papers*, 2015.

[9] A. Abadie, A. Diamond, and J. Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 493–505, 2012.

[10] S. Athey and G. W. Imbens, "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 27, pp. 7353–7360, 2016.

[11] S. Wager and S. Athey, "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," "https://arxiv.org/pdf/1510.04342v3.pdf", 2016.

[12] K. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. Scott, "Inferring Causal Impact Using Bayesian Structural Time-Series Models," *The Annals of Applied Statistics*, vol. 9, no. 1, pp. 247–274, 2015.

[13] J. W. Taylor and P. E. Sharry, "Short-Term Load Forecasting Methods: An Evaluation Based on European Data," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 2213–2219, 2007.

[14] E. E. Elattar, J. Goulermas, and Q. H. Wu, "Electric Load Forecasting Based on Locally Weighted Support Vector Regression," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 4, 2010.

[15] H. Hippert, C. Pedreira, and R. Souza, "Neural Networks for Short-Term Load Forecasting: A Review and Evaluation," *IEEE Transactions on Power Systems*, vol. 16, no. 1, pp. 44–55, 2002.

[16] P. Mirowski, S. Chen, T. K. Ho, and C.-N. Yu, "Demand Forecasting in Smart Grids," *Bell Labs Technical Journal*, 2014.

[17] F. Chen, J. Dai, B. Wang, S. Sahu, M. Naphade, and C.-T. Lu, "Activity Analysis Based on Low Sample Rate Smart Meters," *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 240–248, 2011.

[18] H. Fei, Y. Kim, S. Sahu, M. Naphade, S. K. Mamidipalli, and J. Hutchinson, "Heat Pump Detection from Coarse Grained Smart Meter Data with Positive and Unlabeled Learning," *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1330–1338, 2013.

[19] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private Memoirs of a Smart Meter," *Proceedings of the 2nd ACM Workshop on Embedded Sensing Sysstems for Energy-Efficiency in Building*, pp. 61–66, 2010.

[20] D. Zhou, M. Balandat, and C. Tomlin, "A Bayesian Perspective on Residential Demand Response Using Smart Meter Data," *54th Allerton Conference on Communication, Control, and Computing*, 2016.

[21] ——, "Residential Demand Response Targeting Using Machine Learning with Observational Data," *55th Conference on Decision and Control*, 2016.

[22] A. J. Bagnall and G. J. Janacek, "Clustering Time Series from ARMA Models with Clipped Data," *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 49–58, 2004.

[23] K. K. Jessoe, D. L. Miller, and D. S. Rapson, "Can High-Frequency Data and Non-Experimental Research Designs Recover Causal Effects?" *Working Paper*, 2015.

[24] D. B. Rubin, "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.

[25] J.-S. Pischke and J. D. Angrist, *Mostly Harmless Econometrics*, 1st ed. Princeton University Press, 2009.

[26] P. R. Rosenbaum and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[27] P. W. Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York, 2009.

[29] L. Breiman, J. Friedman, C. Stone, and R. A. Olshen, "Classification and Regression Trees," *CRC Press*, 1984.

[30] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric Statistical Methods*, 3rd ed. John Wiley & Sons, 2013.

[31] "California Irrigation Management Information System," 2017.

[32] T. Vincenty, "Geodetic Inverse Solution Between Antipodal Points," Tech. Rep., 1975.