Residential Demand Response Targeting Using Machine Learning with Observational Data

Datong Zhou, Maximilian Balandat, and Claire Tomlin

Abstract— **The large-scale deployment of Advanced Metering Infrastructure among residential energy customers has served as a boon for energy systems research relying on granular consumption data. Residential Demand Response aims to utilize the flexibility of consumers to reduce their energy usage during times when the grid is strained. Suitable incentive mechanisms to encourage customers to deviate from their usual behavior have to be implemented to correctly control the bids into the wholesale electricity market as a Demand Response provider. In this paper, we present a framework for short-term load forecasting on an individual user level, and relate non-experimental estimates of Demand Response efficacy (the estimated reduction of consumption during Demand Response events) to the variability of a user's consumption. We apply our framework on a dataset from a residential Demand Response program in the Western United States. Our results suggest that users with more variable consumption patterns are more likely to reduce their consumption compared to users with a more regular consumption behavior.**

I. Introduction

The widespread deployment of Advanced Metering Infrastructure (AMI) has made granular data on the electricity consumption of individual residential electricity customers available on a large scale. Smart meters report the electricity consumption of customers at a high temporal resolution, which enables novel data-centric services. One such service is residential Demand Response (DR), in which a DR provider serves as an interface between individual residential customers and the wholesale electricity market. The economic argument made for DR is that it is believed to improve economic efficiency by providing program participants with a proxy of a price signal [1].

Regulators and market operators in different jurisdictions have been moving towards allowing DR providers to offer capacity directly into wholesale electricity markets [2], [3]. The DR provider incentivizes users to temporarily reduce consumption at certain times, e.g. during periods of high Locational Marginal Prices (LMPs), bundles these reductions, and makes capacity bids into the market. If dispatched, the DR provider has to provide a reduction in energy consumption with respect to a certain baseline, and is rewarded by the LMP at the time of dispatch. In such auction-based market settings, it is crucial for DR providers to be able to make informed bids, as bidding too much capacity might result in a penalty due to failure to meet obligations, and bidding too little would result in a suboptimal revenue. The process of making bids is a complex problem - factors to take into account are, among others, the LMP, which determines the marginal price for DR reductions, the number of responsive DR participants under contract, and some knowledge about the behavior of these participants during DR event periods. The DR provider can improve its bidding strategy and efficiency by modeling the users' consumption behavior during DR and non-DR periods and by targeting households with a high potential reduction during DR hours.

In this paper, we identify such users through a combination of established Machine Learning (ML) methods for short-term load forecasting (STLF), load shape clustering, and non-parametric statistics. STLF is employed to predict the consumption of individual residential customers during regular operation as well as during DR periods. This is used in conjunction with a non-parametric hypothesis test to determine whether, under our modeling assumptions, a reduction of consumption during DR periods can be detected [4]. These reductions serve as non-experimental estimates of participants' willingness to reduce energy consumption during DR periods. We then identify a "dictionary" of consumption patterns by clustering load shapes in order to correlate the variability of an individual's consumption pattern to our non-experimental estimate of their consumption shift. Our results show a positive correlation between the degree of variability and our non-experimental estimates of the reductions. This finding may be used for adaptive targeting of users solely based on historical consumption data.

In the area of STLF, the two main categories of research are statistical time series modeling and techniques relying on predictor functions [5]. The first category makes use of ARMA, ARIMA, and SARIMA models [6], and the second uses classical regression techniques such as Least Squares, Lasso- and Ridge-Regression [7], or a class of modern nonparametric methods in which Support Vector Regression [8], Nearest Neighbors Regression, Neural Networks [9], and fuzzy models [9] have been most extensively studied. Other approaches are based on Principal Component Analysis [6], state-space models such as Kalman-Filtering [10] or exponential smoothing methods (Holt-Winters Method) [11].

Clustering algorithms on residential load shapes have been investigated in [12], [13]. A comparison between common clustering algorithms is performed in [14]. Other methods, e.g. Self-Organizing Maps, are explored in [15].

The contribution of this paper lies in combining methods from STLF, load shape analysis, and non-parametric statistics

Datong Zhou is with the Department of Mechanical Engineering, University of California, Berkeley, USA. datong.zhou@berkeley.edu

Maximilian Balandat and Claire Tomlin are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA. [balandat,tomlin]@eecs.berkeley.edu

This work has been supported in part by the National Science Foundation under CPS:FORCES (CNS-1239166).

to identify more responsive users for DR programs. The remainder of this paper is organized as follows: In Section II, we introduce the data and outline preliminary steps. We describe ML algorithms used for STLF in Section III and detail the estimation of energy reduction during DR hours in Section IV. In Section V, we present the methodology used for load shape analysis, and then apply our methods on both synthetic data (Section VI) and real consumption data (Section VII). We conclude in Section VIII.

II. Preliminaries

A. Data Characteristics

Our analyses are based on hourly smart meter readings of 500 residential electricity customers in the western United States, collected between 2012 and 2014. Aligned with those readings are timestamps of notifications sent by the DR provider to the users that prompt them to reduce their consumption for a short period. We also use ambient air temperature measurements from public data sources to capture the correlation between temperature and consumption.

B. Data Preprocessing

Before any analysis is carried out, the data is pre-processed to provide a coherent basis for a comparison of different forecasting techniques.

First, we exclude users with residential solar photovoltaics to remove effects due to correlation in power generation and DR events. We also exclude users with corrupt meter readings (such as excessive or negative consumption).

Second, the time series for consumption and temperature are matched by only taking data into account that includes both temperature and consumption readings. Temperature observations are resampled to hourly data by taking a weighted mean between non-evenly spaced measurements.

Third, consumption and temperature are standardized to zero mean and unit variance to allow future comparisons of prediction methods that are not necessarily scale-invariant.

Fourth, the consumption series are analyzed for stationarity with the augmented Dickey-Fuller test [16]. In particular, it has to be asserted that DR events, interpreted as exogenous "shocks", only have transitory effects and thus do not permanently impact the non-DR consumption. After differencing the consumption series in order to free it from seasonality, all the consumption time series are found to be stationary with a significance level of more then 99%. This is in accordance with [11], where the authors KPSS Test to assert stationarity.

Fifth, and most importantly, hours of meter readings "shortly" after DR hours are removed from the training data. For every DR message sent, 8 hours of subsequent metering recordings are removed to prevent forecasting algorithms to learn from hours that have been influenced by users deviating from their usual consumption behavior. We therefore assume that users revert to their usual behavior at most 8 hours after receiving a DR message. Since existing literature on the "rebound effect", which describes the increase of electricity consumption after the end of DR periods, is concerned with the consumption in a single hour after the DR event [17], [18], removing 8 hours is a conservative estimate to remove spillovers of consumption anomalies into the training data.

C. Covariates

We regress consumption on the following covariates:

- Previous hourly consumptions (autoregressive term),
- Previous hourly ambient temperatures,
- A categorical variable of length 48 combining the hour of day with a boolean weekend indicator variable.

D. Data Splitting

The pre-processed data is split into a training set that represents users' "usual" consumption during non-DR hours, and a DR set with consumption during DR events. The outcome/covariate pairs for user i are denoted as $\left(Y_i^0, X_i^0\right)$ and $\left(Y_i^1, X_i^1\right)$ for the training and the DR set, respectively.

III. Forecasting Techniques

We apply the following forecasting methods:

- Ordinary Least Squares Regression (OLS)
- Lasso (L1) and Ridge (L2) Regression
- k Nearest Neighbors Regression (KNN)
- Support Vector Regression (SVR)
- Decision Tree Regression (DT)

Due to space limitations, we omit in-depth descriptions of the forecasting techniques. The reader is referred to the full version of this paper [19] and the references therein, in particular [20].

Each model is trained on $\left(Y_i^0, X_i^0\right)$ and applied to the covariates of the DR data X_i^1 to obtain the estimated consumption $\hat{Y}_i^c$. This prediction is then compared to the observed consumption Y_i^1 during DR events. The differences $Y_i^\Delta = Y_i^1 - \hat{Y}_i^c$ will be used to compare the statistical differences between consumption predictions outside and during DR periods.

For benchmarking purposes, we also use a baseline (BL) measure commonly employed by Independent System Operators [21]. We chose the so-called "10 in 10" methodology as defined by the California Independent System Operator, which calculates the BL for a given hour by averaging the consumption of the same hours of past days, excluding DR events. Further, the baseline on a day of a DR event is modified with a so-called *Load Point Adjustment* [22].

IV. Non-Experimental Estimates of DR Treatment Effects

A. Counterfactual DR Consumption

Following the general idea of [4], we use the different models fitted on the training data to obtain a non-experimental estimate of the *counterfactual* consumption $\hat{Y}_i^c$, which is the *consumption during DR times in the hypothetical absence of a DR event*. This consumption certainly cannot be observed, since at all DR times, the DR event has affected the consumption of a given user. This general problem has been referred to as the fundamental problem of causal inference [23]. Since model misspecification cannot be ruled out, any true causal estimate of treatment effects will require

the comparison of different groups in a randomized controlled trial. Since conducting such an experiment involves significant preparation time and cost, the contribution of our approach is that it allows to generate meaningful non-experimental estimates in a much broader range of settings.

As a proxy for the unobservable counterfactual consumption in the absence of a DR event, we use the prediction $\hat{Y}_i^c$ obtained by the cross-validated forecasting techniques presented earlier. We define the average empirical reduction $\hat{\Delta}_i$ for user i during DR hours as

$$\hat{\Delta}_i = \frac{1}{N}\sum_{j=1}^{N}\big(\hat{Y}_i^c(j) - Y_i^1(j)\big), \tag{1}$$

which is simply the sample mean of the componentwise difference between the estimated counterfactual and the actual, observed DR consumption. N represents the number of DR events. The intuition is that the forecasting models have been trained on non-DR data $\big(Y_i^0, X_i^0\big)$, and predictions for DR consumptions $\hat{Y}_i^c$ assume the absence of DR events. Thus, if the mean of the estimated counterfactual consumption exceeds the mean of the actual DR consumption Y_i^1, then, assuming the absence of model mismatch, the difference in means can be interpreted as the mean reduction during DR events. Note that $\hat{\Delta}_i$ is not restricted to positive values.

Equation (1) is an absolute measure that ignores the respective overall consumption level. For a potentially more meaningful, relative measure, we define the *weighted mean percentage reduction* (MPR)

$$\text{MPR} = \frac{1}{N}\sum_{j=1}^{N}\frac{Y_i^1(j) - \hat{Y}_i^c(j)}{|\hat{Y}_i^c(j)|}\cdot 100\%, \tag{2}$$

which normalizes the componentwise deviations by the estimated counterfactual consumption. MPR < 0 corresponds to an estimated average DR reduction of $|\text{MPR}|\%$. Note that a disadvantage of MPR lies in the normalization of the componentwise deviations by $|\hat{Y}_i^c(j)|$, which gives disproportionate errors for small $|\hat{Y}_i^c(j)|$.

B. Nonparametric Hypothesis Test

$\hat{\Delta}_i$ and MPR can be evaluated on a set of DR events on the individual user level to estimate individual treatment effects, or an aggregation of users to estimate average treatment effects. Clearly, the accuracy of the estimated average treatment effects scales with the size of the user base (modulo potentially unmodeled effects).

However, $\hat{\Delta}_i$ and MPR on an individual user level will typically be very noisy due to the volatility of the consumption behavior of a single user. Therefore, we make use of a nonparametric hypothesis test to compare our estimates on individual users, following the approach presented in [4]. This is done by comparing the samples Y_i^1 and $\hat{Y}_i^c$ with the Wilcoxon Signed Rank Test, whose goal is to determine whether these samples stem from different distributions. The null hypothesis is that both samples are generated by the same (unknown) distribution $F(u)$:

$$H_0 : Y_i^1, \hat{Y}_i^c \sim F(u) \Rightarrow \mathbb{E}\big[Y_i^1 - \hat{Y}_i^c\big] = 0. \tag{3}$$

The null hypothesis (3) is juxtaposed with the one-sided alternative hypothesis H_1, which states the existence of a location parameter shift Δ between the data-generating distributions Y_i^1 and $\hat{Y}_i^c$, which are of the same shape:

$$H_1 : Y_i^1 \sim F(u),\ \hat{Y}_i^c \sim F(u)+\Delta,\ \mathbb{E}\big[\hat{Y}_i^c - Y_i^1\big] = \Delta. \tag{4}$$

If H_1 is accepted, this suggests that, within the constraints of our model, the predicted counterfactuals $\hat{Y}_i^c$ are on average greater than the observed DR consumptions Y_i^1, which can be interpreted as a mean reduction by Δ during DR hours. Further, the p-value of the hypothesis test is the probability of making the observations under the null hypothesis.

C. Wilcoxon Signed Rank Test

We use the Wilcoxon Signed Rank Test (WSRT), also called Hodges-Lehmann Estimator, with paired samples $(\hat{Y}_i^c, Y_i^1)$ to compute the p-value and an estimate $\hat{\Delta}$ of the location parameter shift. The latter corresponds to the mean empirical reduction of consumption during DR-events of user i based on the samples $(\hat{Y}_i^c, Y_i^1)$ [24].

V. Segmentation of Users

The consumption behavior of residential electricity customers is highly variable across the population, and many analyses have been performed on the relationship between socioeconomic factors and household energy consumption, e.g. in [25]. Inspired by these approaches, we explore the existence of a relationship between the *variability* of user consumption and our non-experimental estimates of the change in consumption during DR periods. Any conclusion drawn from this analysis would be useful for the purpose of targeting particular consumers and allow a more efficient identification and recruiting of users for DR programs.

A. Load Shape Analysis

The idea is to find a reduced set of representative, "signature" load shapes that describes the consumption patterns observed among all observed load shapes. Following [12], we define a load shape $s(t)$ of 24 hourly values as

$$a = \sum_{t=1}^{24} l(t) \text{ and } s(t) = \frac{l(t)}{a}, \tag{5}$$

where $l(t) \in \mathbb{R}^{24}$ is a daily consumption profile. We only collect weekday consumption patterns, as there is an increased variability of energy consumption during weekends [12]. Next, to reduce the noise stemming from individual daily load shapes, for each user, 5 consecutive weekday load shapes are averaged and treated as a single one. Denote the collection of all 5-day average loads as $\mathcal{S}$. Finding $C_1, \ldots, C_k$ that minimize the squared error

$$\text{SE } = \sum_{s_i \in \mathcal{S}} \left(C_i - s_i\right)^2, \tag{6}$$

where C_i denotes the cluster center closest to a given load shape s_i, is a clustering algorithm with k clusters to be set. Unlike [12], where the authors make use of a two-step,

hierarchical k-means algorithm, we choose the standard k-means algorithm with different values of the number of a-priori defined cluster centers k.

B. Variability of User Consumption

After the k cluster centers have been found, we characterize the *variability* of a given user using the following metrics:

1) Entropy: Each daily load shape of user j is matched to its closest cluster center. Define $p_j(C_i)$ as the frequency count of the event that a daily load shape is matched to centroid i divided by the total number of load shapes. Then the entropy H_j of user j is

$$H_j = -\sum_{i=1}^{k} p_j(C_i)\log(p_j(C_i)). \tag{7}$$

The entropy is minimal $(=0)$ if the user follows a single centroid, and maximal $(=\log(k))$ if all cluster centers are of equal occurrence [12].

2) Hourly Standard Deviations: We suggest the metric

$$\tilde{H}_j = \sum_{i=1}^{24} \text{std}\left[s_j(i)\right], \tag{8}$$

i.e. the sum of the standard deviations of the observed hourly consumptions over all hours for a given user j. This method has the advantage that it avoids the need to a-priori define the number of clusters k.

VI. Validation on Synthetic Data

We now construct synthetic data to verify the functionality of our forecasting algorithms and predicted counterfactual consumption to estimate the DR reduction. Our motivation is that we can benchmark our models on the a-priori known ground truth of the synthetic data. The goal is to show that, within the limitations of our model, our learning algorithms are capable of predicting the average empirical reduction (1) and the MPR (2) with acceptable accuracy.

To generate an artificial time series $\bar{l}(t)$, a base consumption consisting of the daily characteristic load shapes shown in Figure 5 is constructed. The relative occurrence of the 12 dictionary load shapes in the base consumption is varied so as to generate time series with different entropies (7). Next, a linear temperature contribution as well as Gaussian Noise $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ are added. Further, a random subset of the time indices are defined as DR hours, for which the respective consumption is decreased by a constant $c_{\text{DR}} > 0$. The resulting artifical load shape $\bar{l}(t)$ therefore includes (known) components of the daily characteristic load shapes, the ambient temperature, and DR reductions:

$$\bar{l}(t) = C_i(t) + c_t \cdot T(t) - \mathbb{I}(t \in \mathcal{D}) \cdot c_{\text{DR}} + \epsilon(t), \tag{9}$$

where $\mathcal{D}$ denotes the set of DR times, $C_i(t)$ the cluster center in the base consumption at time t, and c_t the proportionality constant for the ambient temperature at time t. After standardizing $\bar{l}(t)$, we apply forecasting techniques on $\bar{l}(t)$ with the same features used in Section II-C and investigate the prediction accuracy as well as the estimates of the DR reductions as a function of entropy and magnitude of noise.

Figure 1 shows scatter plots for three different noise levels σ estimated with Ridge-Regression. The plot shows the differences between actual and predicted MPR (2), the differences between the known mean reduction and the estimated mean reduction (1), the estimated location parameter shift $\hat{\Delta}$ from the WSRT, and the mean absolute percentage error (MAPE, (10)) of the consumption predictions. Subplots 1-2 indicate that higher noise levels do not qualitatively impact the accuracy of prediction for MPR and the empirical reduction, even though the range of errors increases as σ increases. Similarly, the estimated location parameter shift $\hat{\Delta}$ from the WSRT varies around a constant, which, from further analyses, is found to be c_{DR}. As expected, higher noise levels increase the MAPE of the predictions. The observations imply that, under the correct model specification and in the absence of confounding variables, Ridge-Regression is capable of correctly estimating the MPR, the empirical reduction, and the location parameter shift given by the WSRT, even in the presence of noise. Important, the findings of Subplots 1-3 are independent of entropy. Only subplot 4 shows an increase of MAPE as entropy increases, which is consistent with intuition because more variable consumption is inherently harder to predict. Lastly, further analyses show that the qualitative nature of Figure 1 varies with the bias of the estimator, in the sense that upward biased estimates yield a higher $\hat{\Delta}$.

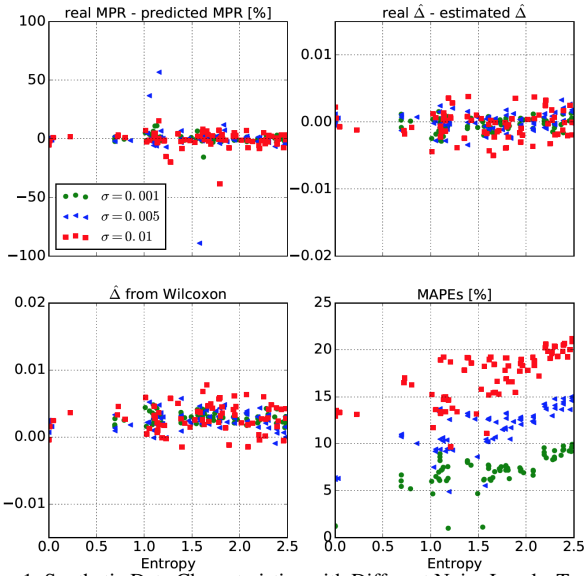

Fig. 1: Synthetic Data Characteristics with Different Noise Levels. Top Left: Actual MPR − Predicted MPR, Top Right: Actual $\hat{\Delta}$ − Predicted $\hat{\Delta}$, Bottom Left: Wilcoxon-$\hat{\Delta}$, Bottom Right: MAPEs

We can think of real load shapes as a mixture of base load shapes, which describe the daily behavior of users. These base loads are then perturbed with temperature influences (e.g. increased AC consumption during hot days) and noise (e.g. user vagaries). It can be imagined that different users possess different archetypes of consumption behavior (e.g. a single person household might have a more regular consumption pattern than a family), and thus different entropies.

Since our analysis on the synthetic data shows that the mean predicted DR reductions are independent of entropy, we conclude that our prediction algorithms are applicable to participants with different levels of consumption variability.

VII. Experiments on Data

A. Prediction Accuracy

We use the mean absolute percentage error (MAPE) as a measure for prediction accuracy:

$$\text{MAPE} = \frac{1}{N}\sum_{j=1}^{N}\frac{|Y_i^0(j) - \hat{Y}_i^0(j)|}{Y_i^0(j)} \cdot 100\%. \qquad (10)$$

Figure 2 shows box plots for the MAPE of different prediction methods and the CAISO baseline across the user population. L1, L2 and LS have similar MAPEs, which indicates that because of the large data set available overfitting is not an issue. As expected, the ISO baseline prediction performs worst since it averages hourly consumption readings far back in the past (up to 10 weekdays before a prediction), which are unlikely to predict the consumption accurately. DTs and SVR with median MAPEs of $\sim$ 23 and 29%, respectively, outperform KNN and the linear regression methods whose median MAPE across users is $\sim$ 30-35 %. However, computation times of up to 45 minutes to fit an SVR model on a time series of length 40,000 were observed, compared to < 5 seconds per user for the linear models (on a six-core CPU). Prediction times for all methods, however, were negligible.

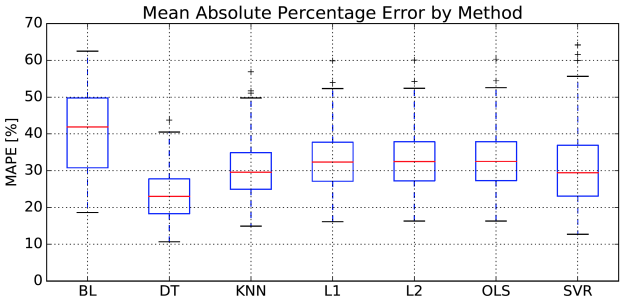

Fig. 2: MAPEs for Different Forecasting Techniques and CAISO Baseline

We acknowledge that more accurate predictions can likely be obtained by taking into account more covariates, e.g. a greater number of autoregressive consumption terms, more temperature data, and more sophisticated ML algorithms such as neural networks. This, however, is not the focus of this paper, and the reader is referred to [9] for a discussion on the performance of forecasting algorithms.

B. Estimation of Reduction of DR Consumption

Figure 3 shows box plots of the estimated treatment effects determined by (1) and estimates of $\hat{\Delta}$ provided by the WSRT, and Figure 4 gives box plots of the range of estimated MPRs across all users by method, computed with (2).

In Figure 3 it can be seen that the median of the mean empirical reductions, computed with both the WSRT and (1), are greater than zero throughout. As already mentioned, the different levels of $\hat{\Delta}$ can be explained by potentially biased estimators, e.g. downward biased estimates of $\hat{Y}_i^c$, on average, yield a smaller $\hat{\Delta}$ [4]. Indeed, our findings reveal that both KNN and SVR yield downward biased estimates across all users with a median value of 0.0025 and 0.0034, respectively. The bias for L1, L2 and DT was found to be less than 10^{-9} for all users. This explains the smaller median $\hat{\Delta}$ for KNN and SVR.

According to Figure 4, the median MPRs are between -0.2% and -7% for all methods except DT, which is synonymous with a DR reduction in all cases but DT. It is seen that the downward biased methods SVR and KNN result in a smaller median reduction. For DT, the counterintuitive result of an increased DR consumption (MPR > 0) despite a near zero bias could possibly be explained with the normalization of some $\left(Y_i^1(j) - \hat{Y}_i^1(j)\right)$ by outliers in $|\hat{Y}_i^1(j)|$ that are close to zero due to misclassifications in the training step.

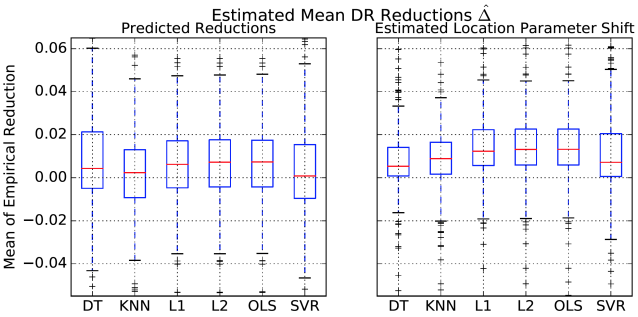

Fig. 3: Estimated DR Reductions. Left: Computed with (1); Right: Wilcoxon-$\hat{\Delta}$

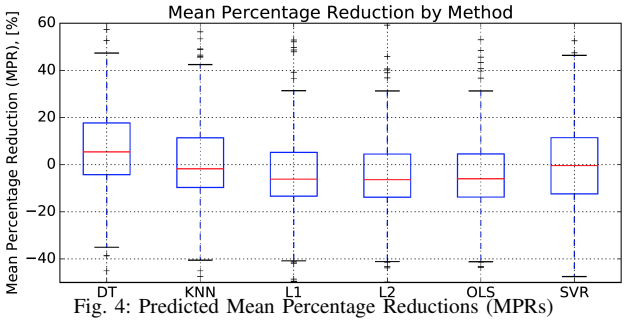

Fig. 4: Predicted Mean Percentage Reductions (MPRs)

C. K-Means Clustering Results

Figure 5 shows the 12 characteristic centroids and the number of load shapes that belong to the respective centroid. Similar to [12], we can characterize different habits of users,

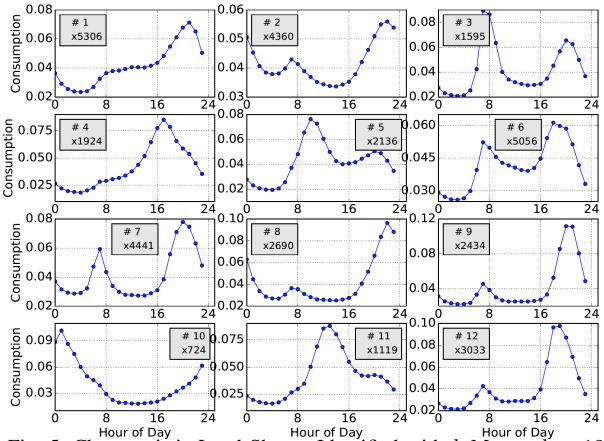

Fig. 5: Characteristic Load Shapes Identified with k-Means, $k = 12$

such as users with a morning and evening peak (#2, #3, #6, #7, #9, #12), daytime peak (#5, #11), night peak (#8, #10), and evening peak (#1, #4).

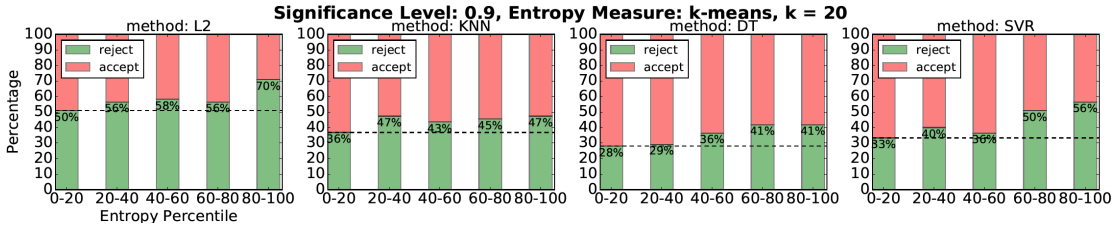

Fig. 6: Percentage of Rejected / Accepted Nulls for k-Means, 20 Clusters and Significance Level $(1-p) = 0.9$

D. Entropies and P-Values

Figure 6 depicts bar charts for the percentage of accepted and rejected Null Hypotheses as defined in (3) for a 90% significance level and different forecasting methods, sorted by entropy percentiles computed with (7) for $k = 20$. Clearly, the percentage of rejections tends to increase as entropy increases. Under the assumption of a correctly specified model and in the absence of confounding variables, this suggests that users with higher variability in their consumption tend to have a lower consumption during DR events than those with lower variability in their consumption. Figure 6 shows a similar trend for different significance levels, k-means with 6 or 12 centroids as well as the standard deviation (8) as entropy criteria. An interesting observation is the tendency towards higher rejection rates for the linear regression models compared to the nonparametric ones. This can be explained by the downward biased estimates of SVR and KNN, which reduce the estimated location parameter shifts $\hat{\Delta}$. A lower estimated location shift results in a smaller test statistic U, which then correlates with fewer rejected nulls in expectation.

VIII. Conclusion

We analyzed Machine Learning methods for predicting residential energy consumption and used them in conjunction with a non-parametric hypothesis test to estimate users' consumption reductions during peak hours. We presented two entropy criteria for the variability of individual household consumption and identified a positive correlation between their inherent variability and the magnitude of the non-experimental estimates of reductions during DR periods.

The covariates used in our approach proved to yield satisfactory prediction results, and an improved choice of training features will only improve the forecasting accuracy, but not change our findings qualitatively. Further improvements can be achieved by incorporating a larger data set with more households and using more refined clustering methods.

The effect of biased forecasts on the estimated DR reductions highlights the need for a more careful evaluation of the employed prediction methods, an issue that we are currently exploring. Due to the non-experimental nature of our estimates, in order to make claims about being able to identify the causal effects of DR interventions, our methods will need to be benchmarked against a randomized experiment.

References

[1] Federal Energy Regulatory Commission (FERC), "National Action Plan on Demand Response," June 2010.

[2] PJM Interconnection LLC, "PJM Capacity Performance Updated Proposal," Oct 2014.

[3] Public Utilities Commission of the State of California (CPUC), "Resolution E-4728. Approval with Modifications to the Joint Utility Proposal for a DR Auction Mechanism Pilot," July 2015.

[4] M. Balandat, D. Zhou, and C. Tomlin, "Machine Learning Methods for Causal Inference on Time Series Data," *In preparation*, 2016.

[5] A. Muñoz, E. F. Sánchez-Úbeda, A. Cruz, and J. Marín, *Handbook of Power Systems II*, S. R. et al., Ed. Springer-Verlag, 2010.

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

[7] S. Fan and R. J. Hyndman, "Short-Term Load Forecasting Based on a Semi-Parametric Additive Model," *IEEE Transactions on Power Systems*, vol. 27, no. 1, February 2012.

[8] 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.

[9] R. E. Edwards, J. New, and L. E. Parker, "Predicting Future Hourly Residential Electrical Consumption: A Machine Learning Case Study," *Energy and Buildings*, vol. 49, pp. 591–603, 2012.

[10] M. Ghofrani, M. Hassanzadeh, M. Etezadi-Amoli, and M. Fadali, "Smart Meter Based Short-Term Load Forecasting for Residential Customers," *North American Power Symposium (NAPS)*, 2011.

[11] P. Mirowski, S. Chen, T. K. Ho, and C.-N. Yu, "Demand Forecasting in Smart Grids," *Bell Labs Technical Journal*, vol. 18, no. 4, 2014.

[12] J. Kwac, J. Flora, and R. Rajagopal, "Household Energy Consumption Segmentation Using Hourly Data," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, 2014.

[13] J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber, "Clustering Analysis of Residential Electricity Demand Profiles," *Applied Energy*, vol. 135, pp. 461–471, 2014.

[14] Y.-I. Kim, J.-M. Ko, and S.-H. Choi, "Methods for Generating TLPs (Typical Load Profiles) for Smart Grid-Based Energy Programs," *Computational Intelligence Applications in Smart Grid (CIASG)*, 2011.

[15] G. Chicco, "Overview and Performance Assessment of the Clustering Methods for Electrical Load Pattern Grouping," *Energy*, vol. 42, pp. 68–80, 2012.

[16] W. A. Fuller, *Introduction to Statistical Time Series*. Wiley-Interscience, 1995.

[17] J. L. Mathieu, D. S. Callaway, and S. Kiliccote, "Examining Uncertainty in DR Baseline Models and Variability in Automated Responses to Dynamic Pricing," *Conference on Decision and Control*, 2011.

[18] Y. Li, B. L. Ng, M. Trayer, and L. Liu, "Automated Residential Demand Response: Algorithmic Implications of Pricing Models," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, 2012.

[19] D. Zhou, M. Balandat, and C. Tomlin. (2016) Residential Demand Response Targeting Using Machine Learning with Observational Data. [Online]. Available: http://arxiv.org/pdf/1607.00595.pdf?

[20] ——, "A Bayesian Perspective on A Bayesian Perspective on Residential Demand Response Using Smart Meter Data," *54th Annual Allerton Conference on Communication, Control, and Computing*, 2016.

[21] K. Inc., "PJM Empirical Analysis of Demand Response Baseline Methods," Apr 2011.

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

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

[24] T. P. Hettmansperger and J. W. McKean, "Robust Nonparametric Statistical Methods," *CRC Press*, 2011.

[25] S. Bhattacharjee and G. Reichard, "Socio-Economic Factors Affecting Individual Household Energy Consumption: A Systematic Review," *5th International Conference on Energy Sustainability*, 2011.