Lawrence Berkeley
National Laboratory

Predicting Baseline for Analysis of Electricity Pricing

Taehoon Kim
Dongeun Lee
Jaesik Choi
Anna Spurlock
Alex Sim
Annika Todd
Kesheng Wu

Lawrence Berkeley National Laboratory
One Cyclotron Road, Berkeley, CA 94720

Predicting Baseline for Analysis of Electricity Pricing

Taehoon Kim[1], Dongeun Lee[1], Jaesik Choi[1], Anna Spurlock[2], Alex Sim[2], Annika Todd[2], and Kesheng Wu[2]

[2]Lawrence Berkeley National Laboratory, Berkeley CA, USA
[1]Ulsan National Institute of Science and Technology, Ulsan, Korea

March 30, 2016

Abstract

To understand the impact of new pricing structure on residential electricity demands, we need a baseline model that captures every factor other than the new price. The standard baseline is a randomized control group, however, a good control group is hard to design. This motivates us to devlop data-driven approaches. We explored many techniques and designed a strategy, named LTAP, that could predict the hourly usage years ahead. The key challenge in this process is that the daily cycle of electricity demand peaks a few hours after the temperature reaching its peak. Existing methods rely on the lagged variables of recent past usages to enforce this daily cycle. These methods have trouble making predictions years ahead. LTAP avoids this trouble by assuming the daily usage profile is determined by temperature and other factors. In a comparison against a well-designed control group, LTAP is found to produce accurate predictions.

1 Introduction

With measurements recorded for most customers in a service territory at hourly or more frequent intervals, advanced metering infrastructure (AMI) captures electricity consumption in unprecedented spatial and temporal detail. This vast and fast growing stream of data, together with cutting-edge data science techniques and behavioral theories, enables behavior analytics: novel insights into patterns of electricity consumption and their underlying drivers [Costa and Kahn, 2013, Todd et al., 2014].

As electricity cannot be easily stored, electricity generation must match consumption. When the demand exceeds the generation capacity, a blackout would occur, typically during the time when consumers need electricity the most [Joskow, 2001, Wolak, 2003]. Because increasing generation capacity is expensive and requires years of time to implement, regulators and utility companies have devised a number of pricing schemes intended to discourage unnecessary consumption during peak demand periods.

To measure the effectiveness of a pricing policy on the peak demand, one can analyze electricity usage data generated from AMI. Our work focuses on extracting *baseline models* of household electricity usage for a behavior analytics study [Cappers et al., 2013, Costa and Kahn, 2013, Todd et al., 2014]. The baseline models would ideally capture the pattern of household electricity usage including all feastures except the new pricing schemes. There are numerous challenges in establishing such a model. For example, there are many features that could affect the usage of electricity, and many of these features, such as the purchase of

new equipement, is information not available to us. Other features, such as outdoor temperature, are known; but their impact is difficult to capture in simple functions.

Although this work shares some similarities with works on forecasting electricity demands and prices [Suganthi and Samuel, 2012, Bianco et al., 2009, Taylor and McSharry, 2007], there are a number of important differences. The fundamental difference between a baseline model and a forecast model is that the baseline model needs to capture the core behavior that persist for a long time, while the forecast model typically aims to forecast for the next few cycles of the time series in question. Typically, techniques that make forecasts for years into the future are based on highly aggregated time series with month or year as time steps [Alfares and Nazeeruddin, 2002, Bianco et al., 2009], whereas those that work on time series with shorter time steps typically focus on making forecasts for the next day or the next few hours [Cottet and Smith, 2003, Oldewurtel et al., 2010, Panagiotelis and Smith, 2008, Taylor, 2010].

In the specific case that has motivated our work, the overall objective is to study the impacts of pricing policies. The process of designing these pricing schemes, recruiting participants for a pilot study, implementing the pricing schemes, and monitoring the impacts have taken a few years. The baseline model is based on observed consumption prior to the implementation of the new pricing schemes, and applied to predict what consumer behavior would have been without the pricing changes. This is challenging because the baseline model not only captures intraday electricity usage but also needs to be applicable for years. Furthermore, in preliminary tests, we have noticed that the impact of the pricing schemes is weaker than the impact of other factors such as temperature, therefore, the baseline model must be able to incorporate the outdoor temperature, which has a complex relationship with the electricity demand.

This work examines a number of methods for developing the baseline models that could satisfy the above requirements. We use a large set of AMI data to exercise these methods and evaluate their relative strengths. The bulk of data in this work is hourly electricity usage from randomly chosen samples of households from a region of the US where the electricity usage is highest in the afternoon and evening during the months of May through August. The current work extracts the baseline models for average behavior of different customer groups, not behavior specific to any individual household.

In the remainder of this paper, we briefly present the background and related work in Section 2 and describe the residential electricity usage data used in this study in Section 3. We also present some analysis with conventional statistical methods in Section 3. We describe the methods used to extract the new type baseline in Section 4 and discuss the output from these methods in Sections 5 and 6. A short summary is provided in Section 7.

2 Application Driver

Energy management has become an important problem all around the world. The recent deployment of residential AMI makes hourly electricity consumption data available for research, which offers a unique opportunity to understand the electricity usage patterns of households. In particular, understanding how and when households use electricity is essential to regulators for increasing the efficiency of power distribution networks and enabling appropriate electricity pricing. One concrete objective from several current pricing studies is to design new rules and structures to reduce the peak demand and therefore level out total electricity usage [Espey and Espey, 2004, Todd et al., 2014].

The influx of massive amounts of electricity data from AMI has led to a variety of research on energy behavior such as electricity consumption segmentation [Chicco et al., 2004, Figueiredo et al., 2005, Verdú et al., 2006, Chicco et al., 2006, Tsekouras et al., 2007, Smith et al., 2012, Kwac et al., 2014], forecasting and load profiling [Espinoza et al., 2005, Irwin et al., 1986, Flath et al., 2012], and targeting customers for

an air-conditioning demand response program to maximize the likelihood of savings [Kwac and Rajagopal, 2013].

An important tool for this problem is classifying and representing different households with different load profiles [Capasso et al., 1994, Flath et al., 2012, Kwac et al., 2014]. Accurately identifying the load profiles will allow the researchers to associate observed electricity usage with consumer energy behavior. Load profiling could identify policy relevant energy lifestyle segmentation strategies, which can lead to better energy policy, improve program effectiveness, increase the accuracy of load forecasting, and create better program evaluation methods [Kwac et al., 2014].

Accurate prediction or load forecasting of electricity usage is very important for the industry [Nogales et al., 2002, Ramchurn et al., 2012]. For example, long-term usage forecasting for more than one year ahead is important for capacity planning and infrastructure investments. Short-term forecasting is used in the day-ahead electricity market, determining available demand response, and increasing demand side flexibility. We can broadly divide these forecasting techniques into black-box techniques and white-box techniques. The black-box approaches focus on what could be extracted from data, typically based on statistical and machine learning methods [Alfares and Nazeeruddin, 2002, Edwards et al., 2012, Espinoza et al., 2005, Irwin et al., 1986, Nogales et al., 2002, Ramchurn et al., 2012, Swan and Ugursal, 2009]. For example, some authors prefer supervised machine learning methods such as support vector machines [Chen et al., 2004, Humeau et al., 2013], some use statistical models such as dynamic regression [Nogales et al., 2002], while others advocate for neural networks and artificial intelligence approaches [Ramchurn et al., 2012]. Typically, these methods transform the time series of historical data into a time scale such that the predictions are made for the next time step or the next few time steps.

White-box approaches are typically based on some understanding of the relationship between some cause and its direct effect. For example, becuase increased outdoor temperature leads to increased indoor temperature, which in turn leads people to turn on their airconditioners, one might come up with a model relating outdoor temperature and electricity usage, and then try to fit the parameters of the model using the observed data. However, such a model most likely would not be able to capture all relevant features, because some of the features, such as length of the day, have weak or unclear effect on electricity usage, and others, such as number of occupants in the building, clearly affect the eletricity usage but their values are unknown or their impact on electricity usage is multifacted or unknown [Borgeson, 2014, Fels, 1986, Rabl and Rialhe, 1992]. For this reason, many researchers refer to these models as "gray-box" models because these models always contain a certain amount of unexplained features left as "errors."

Household electricity usage depends on many features beyond what was mentioned above, for example, appliances in the house, the energy behavior of the occupants, the time of day, day of the week, seasons, and so on [Cappers et al., 2013, Todd et al., 2012]. Some of the existing prediction models focus on aggregated demand and therefore could parameterize many factors affecting the usage of an individual household [Swan and Ugursal, 2009]. From the study of earlier models, we learned that a household's electricity usage is strongly periodic, in that the daily electricity usage repeats every day and every week. Given any two consecutive days, their usage patterns are very similar to each other; given any two consecutive weeks, their electricity uses are also similar to each other. Throughout a year, the overall electricity usage follows the pattern of seasonal temperature change. To accurately predict electricity usage, we need to capture all these factors in our own models.

3 Dataset

Our electricity usage data was collected through a well-designed randomized control trial [Cappers et al., 2013]. It has hourly electricity consumption records of individual households for three years. The unit of electricity is in kilowatt-hour (KWh). The total number of hourly data points is 160,125,432, from which we focus on data generated during the summers, which accounts for most of electricity usage (from June 1 to August 31), yielding 41,698,080 data records. The data records from three years are labelled by $(T-1, T, T+1)$, where year $T-1$ corresponds to the year when the electricity has a fixed price throughout the day, and the new prices are used in year T and $T+1$.

3.1 Groups

The households involved in this study are divided into a number of different groups, in this work, we only use three of them, the Control group, the Passive group and the Active group. Following the general design of a randomized controll trial, the Control group is a random selected set of households that are meant to be used as the baseline [Costa and Kahn, 2013, Concato et al., 2000]. In later discussion, this group is labelled as *Control*. This control group is unaware of the study and stays with the previously available fixed-price scheme throughout the testing period*. The other two groups are generally referred to as the treatment group.

The treatment groups use a time-based price, where during the peak-usage hours, 3PM to 7PM in the region of this study, the per KWh charge is higher than the rest of the day. In the Active group, households have to opt in to the new pricing scheme offered. While the households in the Passive group are informed of their participatioin in the new price trial and offered a chance to opt out of the trial.

As in a typically consumer bahvior study, the response rate of the households invited to participate in the new price trial, only a small fraction of the invitees actually opted in. To avoid the imbalance among the three groups, we randomly selected about 1600 households from each of the three groups. We dropped households that do not have measurement data for the whole duration of the study. The number of households dropped is relatively small.

3.2 Overall statistics

Fig. 1 shows the average daily electricity usages of three groups over three summer seasons. The data from each of the three years are plotted as a separate line. We note that even though different pricing schemes are used, the impact of the pricing schemes is not obvious. This can be partially explained by Fig. 2, where average hour temperatures and electricity usages are plotted against hour.

In Fig. 2, the temperatures of T and $T+1$ are higher than the temperature of $T-1$, which means households have experienced hotter summers in T and $T+1$. As a result, the electricity usage increases in T and $T+1$. Even though the new pricing schemes are designed to reduce electricity usage, but the increases in temperature complicates the analysis. Furthermore, the impact of temperature on electricity usage does not appear to be instantaneous; but its impact on electricity usage appears a few hours later. The increased electricity usage during the summer afternoon is mostly from airconditioning, which is more directly related to the indoor temperature, while the temperature reported in our dataset is outdoor temperature. It takes time

*There was an adjustments of the actual prices of the fixed-price scheme. The standard fixed-price scheme typically has a base charge per month and an additional per KWh charge based on the actual usage, where this per KWh charge is generally known as the rate. Early in year T+1, before the summer, there was an increase in the base charge and descrease in the rate. This price change might encourage households to consume more electricity since the increamental cost has gone down.

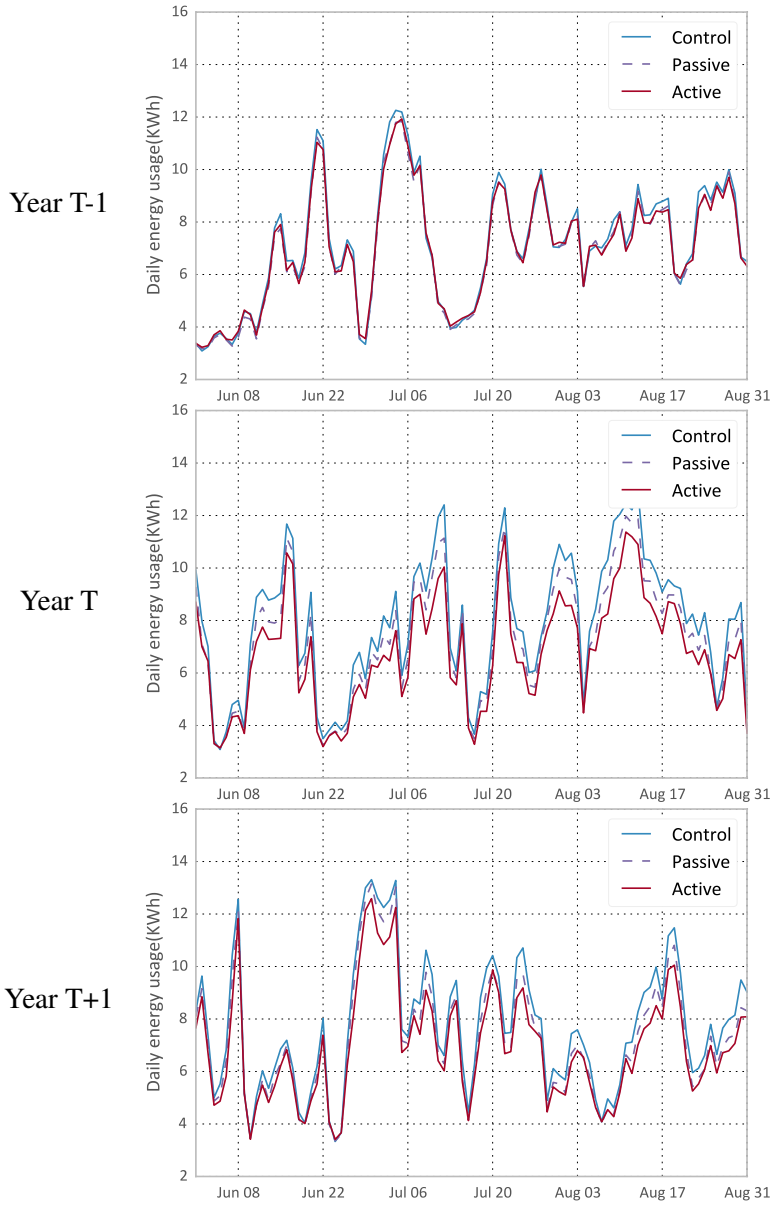

Figure 1: Daily electricity usages of three groups for year $(T-1, T, T+1)$.

for the increased outdoor temperature to impact the indoor temperature. Additionally, residents of a house typically return from work in late afternoon, which increase the number of occupants in a household.

Because there is no obvious differences from Figs. 1 and 2, we conclude that the influence of common features such as season, outdoor temperature, day of the week and so on are much stronger than the features that distinguish the groups. This means the baseline models have to be very accurate in order to recognize the different groups. We will discuss these methods carefully in Section 4.

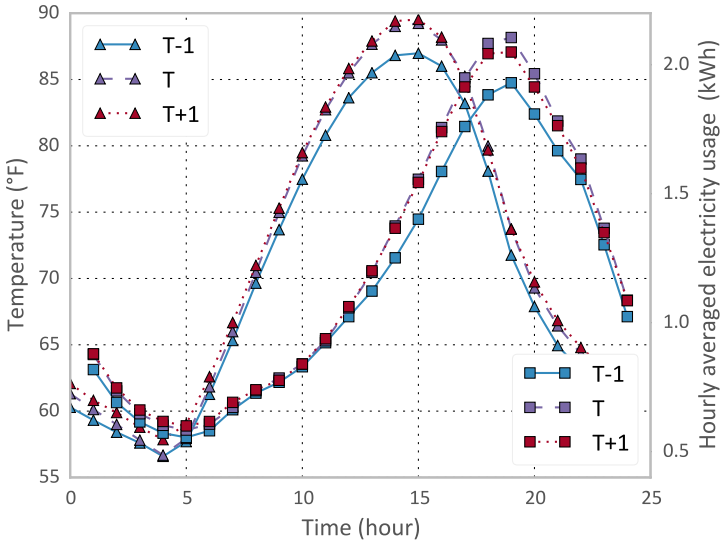

Figure 2: Hourly temperatures (triangle markers) and electricity usages (square markers) for $(T-1, T, T+1)$. Note the time lags between the peaks of temperatures and the peaks of electricity usages, which should be taken into consideration when we express a baseline usage model with outdoor temperatures. The temperatures of T and $T+1$ are higher than that of $T-1$, which results in the higher electricity usages in T and $T+1$.

Table 1: The hourly electricity usages for three groups averaged over all hours of the summer days in each year, and their differences relative to the control group. The values in **bold** are expected to be less than 0.021 in absolute value.

	Average hourly usage			Subtract control		
year	T-1	T	T+1	T-1	T	T+1
Control	1.128	1.205	1.197			
Passive	1.100	1.152	1.154	**-0.028**	-0.053	-0.043
Active	1.125	1.160	1.173	**-0.003**	-0.045	-0.024

Table 2: The hourly electricity usages for three groups averaged over the peak-demand hours of the summer days in each year, and their differences relative to the Control group. In later discussions, these average usage values measured by the smart meters are referred to as M_{T-1}, M_T, and M_{T+1}.

	Average hourly usage			Subtract control		
year	T-1	T	T+1	T-1	T	T+1
Control	1.790	1.973	1.937			
Passive	1.742	1.822	1.818	-0.048	-0.151	-0.119
Active	1.752	1.696	1.739	-0.038	-0.277	-0.198

3.3 Comparison against the control group

In the tradition of randomized controlled trials, our dataset contains a control group. This control group is a valid counterfactual group and can provide a baseline for group-wise comparisons using a Randomized Encouragement Design (RED) evaluation methodology [Todd et al., 2012]. However, we are interested in developing a new baseline methodology that does not rely on a randomized control group [Horwitz and Feinstein, 1979, Liddle et al., 1996]. We are interested in developing such a methodology for two reasons: (i) we would eventually like to use our technique to build a baseline for each household individually, which necessitates the development of new baseline models that do not rely on a control group counterfactual; (ii) it is often the case that programs, such as the pricing programs used in this paper, are implemented by electricity providers without a randomized evaluation methodology. It is often the case that randomization is either impractical, too expensive, or hampered by regulatory requirements. For this reason, it is extremely valuable to have a methodology that can be used to evaluate program effectiveness without relying on randomization. Therefore, we will be using this dataset in order to demonstrate such a methodology. We will use the control group as a comparison group in order to validate the baseline methodology we develop, but will use only the households in the treatment group that self-selected into treatment. If these households were compared directly to the control group, one would be concerned about self-selection bias. Using an accurate baseline methodology is one potential way to avoid such a bias, by allowing for the estimation of the effect of the pricing scheme within those households that self-selected into the study.

Looking first at the broad changes in consumption across the groups. Tables 1 and 2 contain the average hourly electricity consumption for all hours of a day and peak-demand hours, respectively. The values in Table 1 is averaged over all hours and all days of the summer months in each year, while the values in Table 2 is averaged over the peak-demand hours of each summer day. From these numbers, we see that the average hourly usages are higher in year T and year T+1. However, the increases of the two treatment groups are smaller than that of the control group. Relative to the control group, the treatment groups have reduced electricity consumption. This is particularly true during the peak-demand hours as shown in Table 2. These observed changes match the design goal of the new pricing schemes.

In order to underline why a baseline method such as the one we develop is needed, we show here the extent of the self-selection bias that exists if one were simply to compare the self-selected treatment households to the control households. To do this we examine if the differences in year T-1 (before the introduction of the treatments) are within the expected confidence intervals.

The standard deviations of hourly usage values for all households are all about 0.85 (KWh)[†] and each of the group has about 1600 households, therefore, we expect the confidence interval of the these average values to be about $0.85/\sqrt{1600} = 0.021$. For a control group to be considered as properly selected, the differences between the various groups before the introduction of the treatments should be less than 0.021, however among the two relevant difference values in year T-1 only one has a absolute value less than 0.021 in Table 1. This suggests that the three groups are not well randomized, and self-selection bias of the treatment groups could be strongly present in the data. We propose that the baseline method we develop is a solution to this problem.

3.4 Differences among the groups

Next we directly compare the time series of the average hourly usage of each group to understand their differences. For this test, we have selected to compare time series with the Kolmogorov-Smirnov test (KS test) [Conover and Conover, 1980]. Given two time series, the KS test measures the distance between their

[†]The actual values are 0.83 for Year T-1, 0.85 for year T, and 0.91 for year T+1.

Table 3: KS test scores for comparing the hourly electricity time series over three summers. When the KS score is larger than 0.05, the two time series are considered as likely to be generated from the same probability density distribution.

	year T-1	year T	year T+1
Control v. Passive	0.09	0.03	0.02
Control v. Active	0.01	0.04	0.04
Passive v. Active	0.09	0.02	0.03

cumulative distribution functions (CDFs) and produces a score between 0 and 1. In many applications, when this score is greater than 0.05, the two input time series are considered as following the same distribution (or loosely, the "same").

Table 3 shows KS test results for each of the three years. In year T-1, where all groups receives the same pricing scheme, we expect the control group to behave similar to the control groups. In terms of KS test scores, we expect all three KS test scores to be greater than 0.05. However, the control group is clearly different from the active group (because the KS test comparing the two time series has a score less than 0.05), even though the difference between average values of these two time series are fairly close to zero as shown in Table 1. Combining the values from these tables, we have plenty of evidences to suspect that the three groups are not well randomized and the self-selection bias might be prevalent.

The KS test scores for year T and year T+1 are all less than 0.05, which indicate that the time series of hourly electricity usages should be considered different. These differences could possibly be extracted and attributed to the price differences and consumer behavior differences.

4 Methodology

The statistics provided in the previous section suggest that the groups in this study might not be well randomized and therefore the control group might not be a good baseline for comparison. This is one motivation for our attempt at developing alternative baseline models. The second motivation for considering alternative baseline models is that we would like to eventually develop a model that is suitable for studying each individual household, but the randomized control group is only a good baseline for the average behavior of a treatment group, not individual households. In this section we first introduce a few black-box approaches and then introduce a white-box approach. The black-box methods are three statistical machine learning methods: linear regression, gradient linear boosting, and gradient tree boosting. The white-box method is named LTAP.

4.1 Linear Regression

One of popular and simple regression models is the linear regression (LR) where a model is represented in the form of linear equations. Multiple LRs can be used to forecast electricity consumption of households [Bianco et al., 2009]. Given a data set $\{y_i, x_{i,1}, ..., x_{i,K}\}_{i=1}^{n}$ of n statistical units, an LR can be represented as follows:

$$\hat{y}_i = \epsilon + \sum_{k=1}^{K} \beta_k x_{i,k} \tag{1}$$

where $\hat{y}_i$ is an estimated value of y_i, β_k is a kth regression coefficient of $x_{i,k}$, and ϵ is a bias.

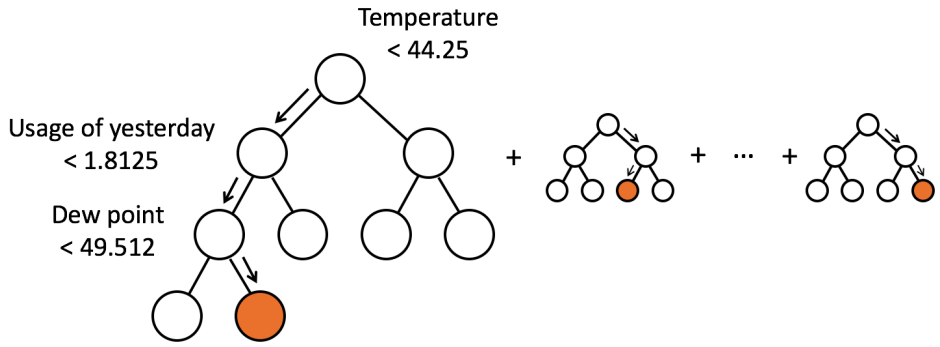

Figure 3: An example of Gradient Tree Boosting (GTB) model. The directed arrow represents a possible path of a sample during the test. Each decision tree decides which path a sample should traverse. Values of leaf nodes are summed to get the prediction.

4.2 Gradient Linear Boosting and Gradient Tree Boosting

Boosting is a prediction algorithm derived from machine learning literature based on the idea of combining a set of weak learners to create a single strong learner. The boosting method has attracted much attention due to its performance on various applications in both machine learning and statistics literature [Schapire, 1990, Freund et al., 1996, Schapire and Freund, 2012].

Gradient Boosting (GB) is one of the boosting methods which constructs an additive regression model by sequentially training weak learners in the gradient descent viewpoint [Friedman, 2001]. GB can be further distinguished by choosing different week leaners. Here we choose two different weak learners: linear function and decision tree. Each model is called Gradient Linear Boosting (GLB) and Gradient Tree Boosting (GTB) respectively.[‡] Fig. 3 shows an example of binary decision trees where each arrow shows a possible path of a sample during testing.

In general, GB can be represented as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}, \tag{2}$$

where K is the number of weak learners, f_k is a function (linear function or decision tree) in the functional space $\mathcal{F}$ which is the set of all possible regression functions, x_i is an input value from a training set, and $\hat{y}_i$ is the estimation of an output value y_i from the training set.

The objective of GB is to minimize the following objective function $obj(\cdot)$ of Θ which denotes the parameters of GB:

$$obj(\Theta) = L(\Theta) + \sum_{k=1}^{K} \Omega(f_k), \tag{3}$$

where $L(\cdot)$ is a training loss function, $\Omega(\cdot)$ is a regularization term. Specifically, we use the root-mean-square error (RMSE) as the training loss function $L(\cdot)$ which is written as:

$$L(\Theta) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \tag{4}$$

[‡]XGBoost library (https://github.com/dmlc/xgboost) is used in this paper.

where n is the number of elements in the training set. We employ hourly training datasets (x_i, y_i) for experiments.

4.3 Linear Relation between Temperature and Aggregated Power (LTAP)

Next, we describe the white-box model that is effective in our tests. It is well-known that the electricity consumption depends on temperature [Fels, 1986]. Generally, this relationship is between the electricity usage of a whole day and the average temperature of that day [Rabl and Rialhe, 1992, Bacher and Madsen, 2011, Borgeson, 2014]. In this work, we propose a simple strategy to make predictions of hourly usage based on this relationship between the daily electricity usage and the average daily temperature. Next, we provide a brief explanation of the rationale for this method before describing the method.

As we see from Figure 2, the relationship between outdoor temperature and the hourly electricity usage is complex, but the daily electricity usage and the average outdoor temperature is relatively straightforward. Since this work is primarily concerned about the peak usages during the summer when airconditioner uses cause the electricity demand to peak in the later afternoon. From the earlier studies on the residential electricity usage, we know there is a significant amount of constant demands from refrigerators, electric water heaters, water pumps, and so on. We assume that this constant usage is the minimum hourly usage during a day and is fixed during the summer season being considered for this work. The usage that is beyond the minimum varies from hour to hour, we call this portion the variable electricity usage. For the region where this data is from, we assume the primary demand for this variable usage is from the airconditioners and therefore is related to the outdoor temperature.

The reason that the daily variable electricity usage is likely a simple function of the average daily temperature can be stated as follows. The higher outdoor temperature causes heat to enter into a house and increases the indoor temperature. When the indoor temperature rises to a certain threshold, the airconditioner starts to cool the room. There is a delay between the rise of outdoor temperature and the rise of the indoor temperature because of the insulation of the house, however, during the warm period of the day, the higher the average temperature causes more heat to enter the house, and more electric power is needed to cool the house. Therefore, we expect the aggregate variable electricity usage per day to have a relatively simple relation with the average outdoor temperature. From the research literature and our own tests presented in the next section, we see that this is true. In fact, we have a set of linear functions relating the aggregate variable electricity usage and the average outdoor temperature. We will use these linear relationships to forecast the total variable electricity usage from the reported outdoor temperature values.

To distribute the aggregate daily usage to hourly usage values, we make the simple assumption that the profile of daily usage per household remains the same, and scale the variable hourly electricity usage proportional to the change in the aggregated usage. Next, we give a more precise definition of the procedure we call LTAP.

Given a summer day in year T or year T+1, we compute the average temperature t_1 of the day from the hour temperature values. Call this the prediction day. Look for a summer day in year T-1 with the closest average temperature t_0. Call this day the reference day. Let the 24 hourly electricity usage be $h_0[i], i = 0, \ldots, 23$. Let $b_0 \equiv \min h_0[i]$ and $a_0 \equiv \sum(h_0[i] - b_0)$. Let s denote the slope of the linear relation between a_0 and t_0. We compute a_1 as follows

$$a_1 = a_0 + s(t_1 - t_0). \tag{5}$$

We assign the hourly electricity usage as follows

$$h_1[i] = b_0 + (h_0[i] - b_0)a_1/a_0. \tag{6}$$

Table 4: RMSE for Three Different Models: Linear Regression (LR), Gradient Linear Boosting (GLB), And Gradient Tree Boosting (GTB).

	LR	GLB	GTB
Control	1.841	0.952	**0.845**
Passive	1.862	0.951	**0.838**
Active	1.731	0.957	**0.839**

It is easy to verify that the above assignment of the aggregated electricity usage to each hour preserves the shape of the daily usage profile while giving the correct total usage value as predicted by Equation 5. Furthermore, this prediction algorithm does not involve any explicit values of days and therefore can be applied to any day.

5 Black-box Regression Models

To establish our baseline, we need to first determine the features that this model depends on. From information in the literature and our exploration of the dataset, we choose 8 features: 3 time variables (month, hour, and day of week), 2 historical electricity usage variable (electricity usage of the same hours on a day before (yesterday) and a week before), and 3 hourly averaged weather conditions (temperature, atmospheric pressure, and dew point). The role of the historical usage data is to distinguish each household from others. Here, the weather data vary only over time, not across households, since all households belong to a geographical region covered by a single weather station. Although some weather data such as the atmospheric pressure and the dew point do not seem to play major roles at first glance, we also want to take them into account to see whether there is a latent correlation between these data and electricity usage.

5.1 Errors of the models

We explore three different models: LR, GLB, and GTB, described in Section 4, and plan to choose a single model that best represents the core behavior. Specifically, we trained the three models with the usage data in $T-1$ by randomly sampling 70% of data as a training set and using the remaining 30% of data as a test set. In the case of GLB and GTB, we trained 1,000 decision trees for a single GTB. If the sum of child nodes' weights was less than 2, we kept partitioning a tree before the max depth of tree surpassed 5. For each step, we randomly collected half of the data set and shrink the feature weights to 0.3 so as to avoid overfitting. These parameters were provided by XGBoost package and we tuned hyper parameters using 5-fold cross-validation with a grid-search method in the parameter spaces.

Table 4 shows the result of RMSE for the three models. We see that the errors of LR, GLB are larger than GTB. This is not unexpected since the relationship between electricity usage and temperature is not only non-linear but also delayed. In this work, we choose GTB to extract the baseline.

5.2 Training Gradient Tree Boosting

Our goal is to predict residential electricity consumption with a model that captures the effect of outdoor temperature, including its delayed effect. To achieve this goal, we trained a GTB model with the usage data

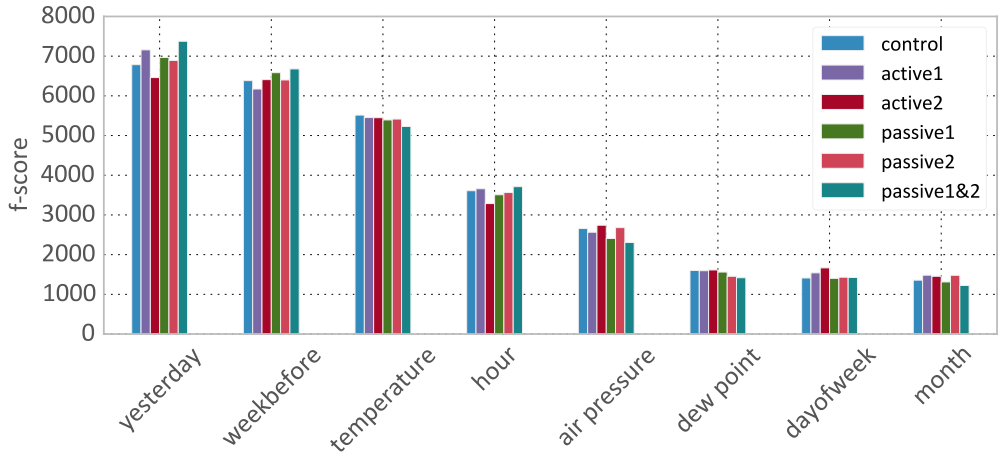

Figure 4: F-score representing the importance of a feature in the decision trees of GBT, which is calculated by counting the appearance of a feature.

of $T-1$ for all households regardless of their groups. Again we randomly sampled 70% of the data as a training set and used the remaining 30% as a test set.

Fig. 4 shows f-score of each feature in GTB, where the f-score is the number of appearances of a feature in all of weak decision trees in GTB. If the f-score of one feature is higher, the feature is more important than other features. The two most powerful features are historical electricity usage data (yesterday and week before usage) and the third most influential feature is temperature. In Fig. 4, we can see how GTB finds which features are important. It is also interesting to note that 'day of week' is not as effective as other features, because we originally assumed that GTB might detect the difference between weekend and weekday from the dataset.

5.3 Hourly Averaged Prediction

Fig. 5 shows the hourly usage prediction by GTB and hourly average temperature of different groups. In year T and $T+1$, we see that the control group uses slightly more electricity than predicted by GTB model, while the treatment groups use less electricity, especially during the peak-demand hours, than the predictions by GTB models. Furthmore, we see that the points representing the measured usages are noticeably blow the lines representing the predictions. Clearly, the new pricing scheme has an impact on the consumer behavior, and the active group has responded more than the passive group. We also see that the GTB model effectively has learned the lagged effect of temperature explained in Fig. 2.

5.4 Modifying GTB for continuous prediction

The features used for our GTB model include the electricity usage from a day ago and a week ago. The current implementation of GTB requires these values to be supplied together with other values that are known beforehand. In the training steps where all the values from year T-1 are considered known values, we should be able to supply the values of these lagged variables as well. However, when making predictions for the future, say for year T, the prediction mechanism is expected to treat electricy usage values as unknown, therefore, the usage values of a day ago and a week ago are only available as more predictions are made. We have modified the GTB prediction procedure to make predictions one day at a time, and use the predicted

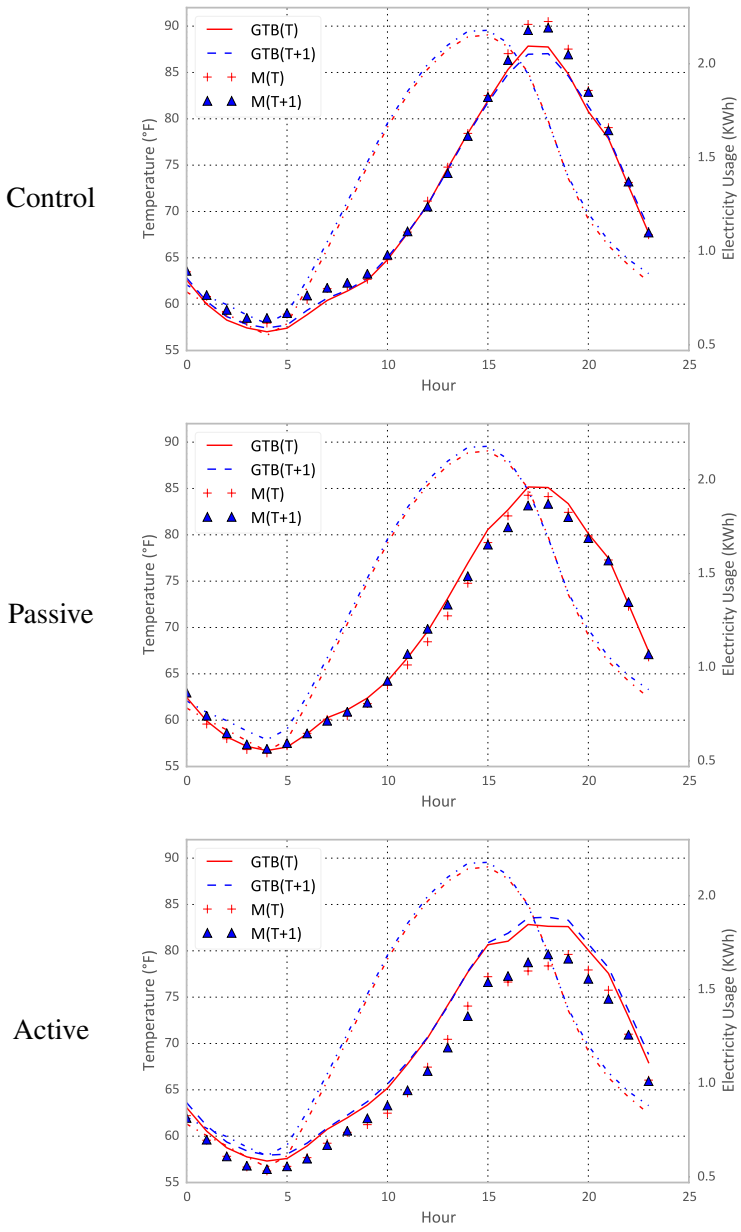

Figure 5: Predicted (by GTB) and measured hourly average electricity usage during year T and T+1. The lines and symbols represent data from the same year have the same color. The measured values are lower than the predictions indicating the consumers have reduced electricity uses compared to the "business-as-usual" predictions.

values for the day ago and week ago usage values. This modified version of prediction procedure as the sequential prediction since it makes predictions one day at a time and immediately make uses of the predicted values.

Fig. 6 shows an attempt to make prediction for a month of time using the above procedure of continuous prediction. We note that as time progresses, the maximum values in each graph gradually increases. This

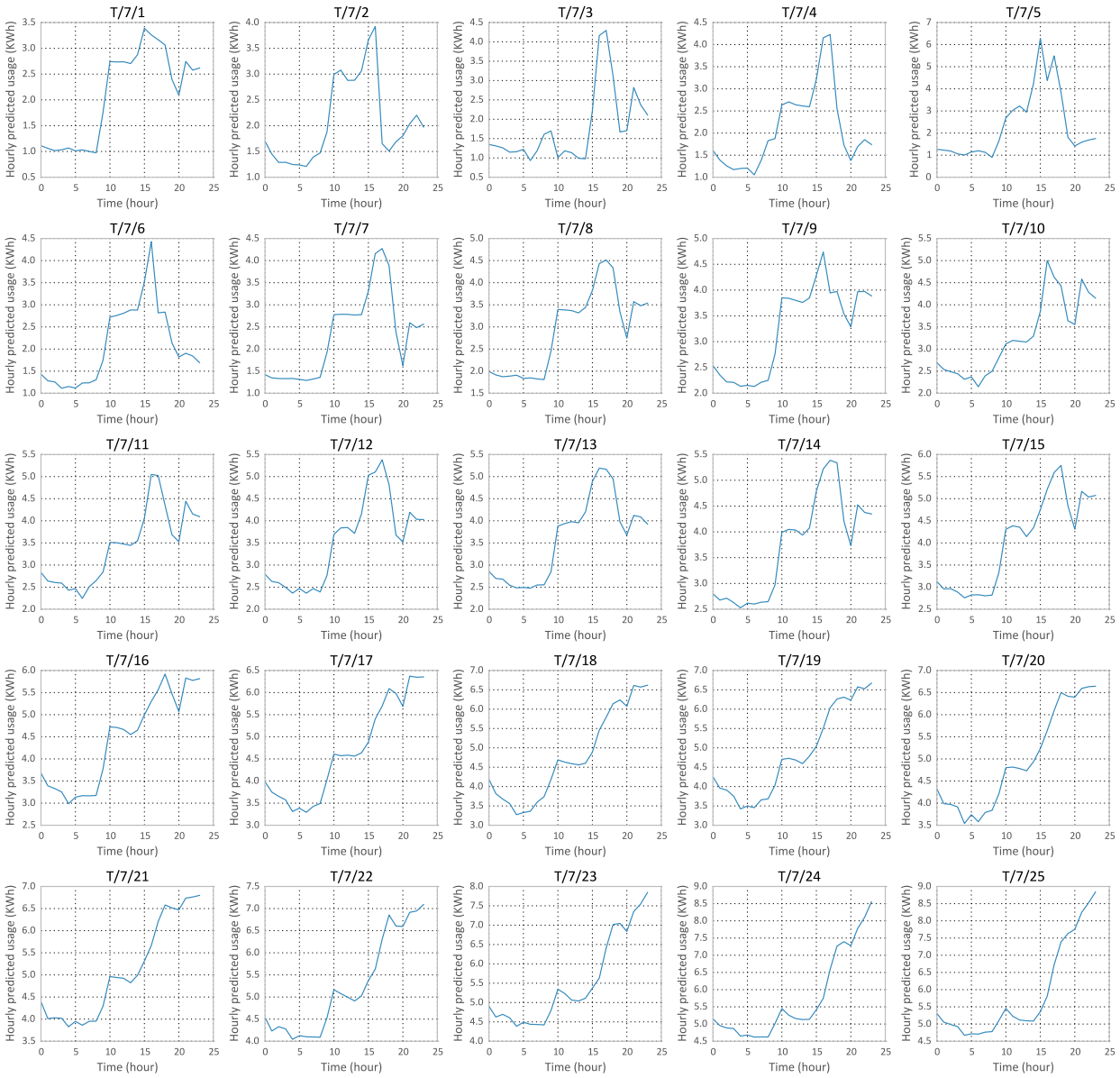

Figure 6: Predicted electricity usage with a modified version of Gradient Tree Boosting (GTB) that uses previous predictions as lagged variables. This is for the 2nd month of the summer, we see the predicted usages are higher than normal at the beginning of the month and continue to grow over time.

appears to be an accumulation of the some sort of prediction errors over time. Typically, predictions are only made for a small number of steps beyond the end of the known time series, however, to establish a baseline for years requires us to make predictions many time steps beyond the end of the known time series. To remedy this problem, we could avoid using lagged variables as features or devise "stable" prediction methods that would not accumulate prediction errors. The LTAP method is a strategy that only makes use of the temperature, and avoid building up new predictions from previous predictions.

Table 5: The slopes and the coefficients of correlation for data points with average temperature above $65°$F from the summer of year T-1.

	slope	coeff of corr
Control	1.13	0.92
Passive	1.07	0.92
Active	1.02	0.91

Table 6: The averaged (over all hours) hourly usage predicted by LTAP and their differences from the actual measurements.

group	D_T	D_{T+1}	$M_T - P_T$	$M_{T+1} - P_{T+1}$
Control	1.185	1.220	0.020	-0.023
Passive	1.156	1.193	-0.004	-0.038
Active	1.181	1.211	-0.021	-0.038

6 White-box Prediction

In Section 4.3, we describe the white-box prediction called LTAP. In this section, we first provide evidence that the assumed linear relationship between the aggregated variable electricity usage and the average daily temperature is valid, and then describe the results of predictions with LTAP.

6.1 Linear relationship between aggregated power usage and temperature

In Section 4.3 we provide some arguments for the a linear relationship between the aggregated variable electricity usage and average daily temperature. Figure 7 and Table 5 provide some empirical support for these arguments. In Figure 7, we provide scatter plots of the aggregated variable electricity usage against the average daily temperature. These scatter plots suggest that below $65°$F, there is no obvious relationship between the electricity usage and temperature, however, at higher temperatures there is clearly a linear relationship between electricity usage and temperature. When more seasons are considered, there are more variety of relationships between electricity and temperature [Rabl and Rialhe, 1992, Bacher and Madsen, 2011, Borgeson, 2014], however, since we are only studying the electricity usage in the summer season of a region where airconditioning is heavily used, it is unsurprising that we observe a simpler relation between temperature and electricity usage.

What is somewhat surprising is that coefficients of correlation in all three groups are above 0.9, which indicates the linear relationship is very strong. Therefore, we should expect this linear function could be used to make accurate predictions about the electricity usage in year T and year T+1.

6.2 LTAP prediction results

The test results in the previous section clearly establishes that the relationship between the aggregated electricity usage and the average temperature to be piece-wise linear, therefore we could attempt to use the LTAP prediction method. This method captures the impact of the temperature, which appears to be the most reliable feature that could be used to make predictions. Other factors we initially suspected to be impactful, such as the day of the week, have found to be less important. At this time, we only use the temperature as

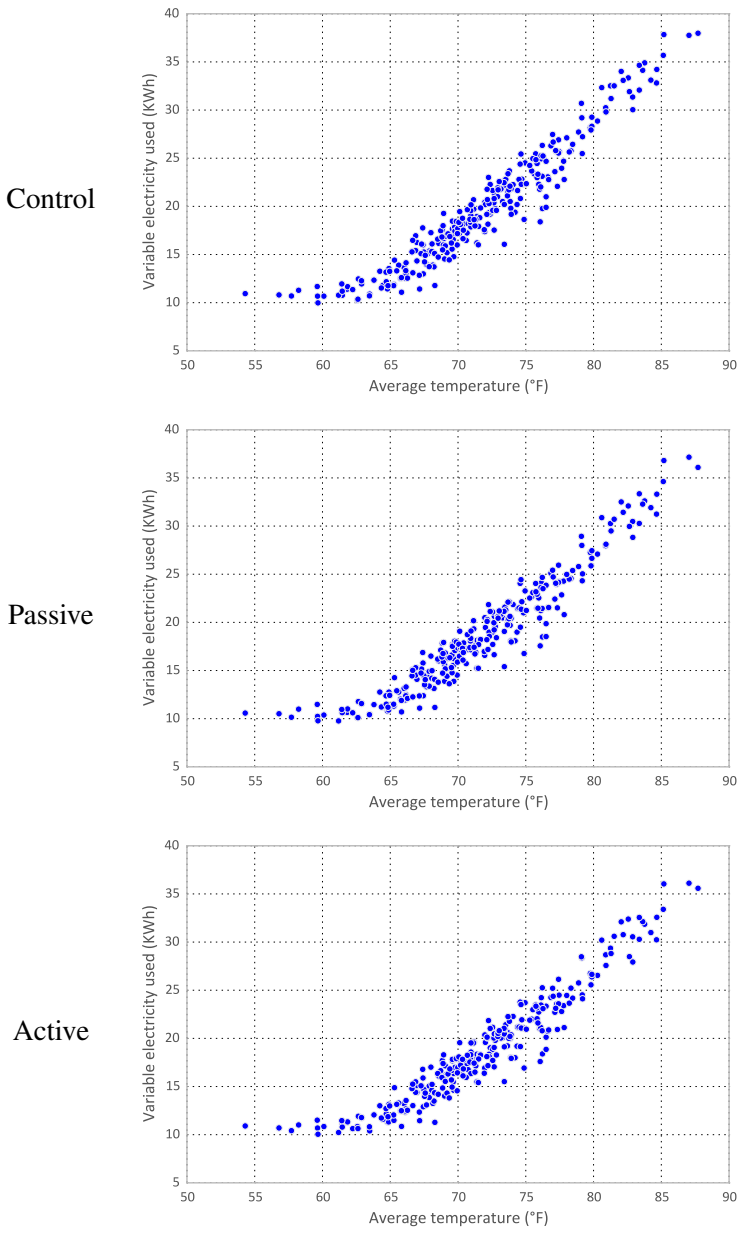

Figure 7: Scatter plots of aggregated variable electricity usage and average daily temperature from the summer of year T-1.

Table 7: The average hourly electricity demand during peak-demand hours and their differences from the actual measurements. The predictions are made with LTAP.

group	P_T	P_{T+1}	$M_T - P_T$	$M_{T+1} - P_{T+1}$
Control	1.960	2.052	0.013	-0.116
passive2	1.904	1.998	-0.081	-0.180
Active	1.910	1.990	-0.214	-0.251

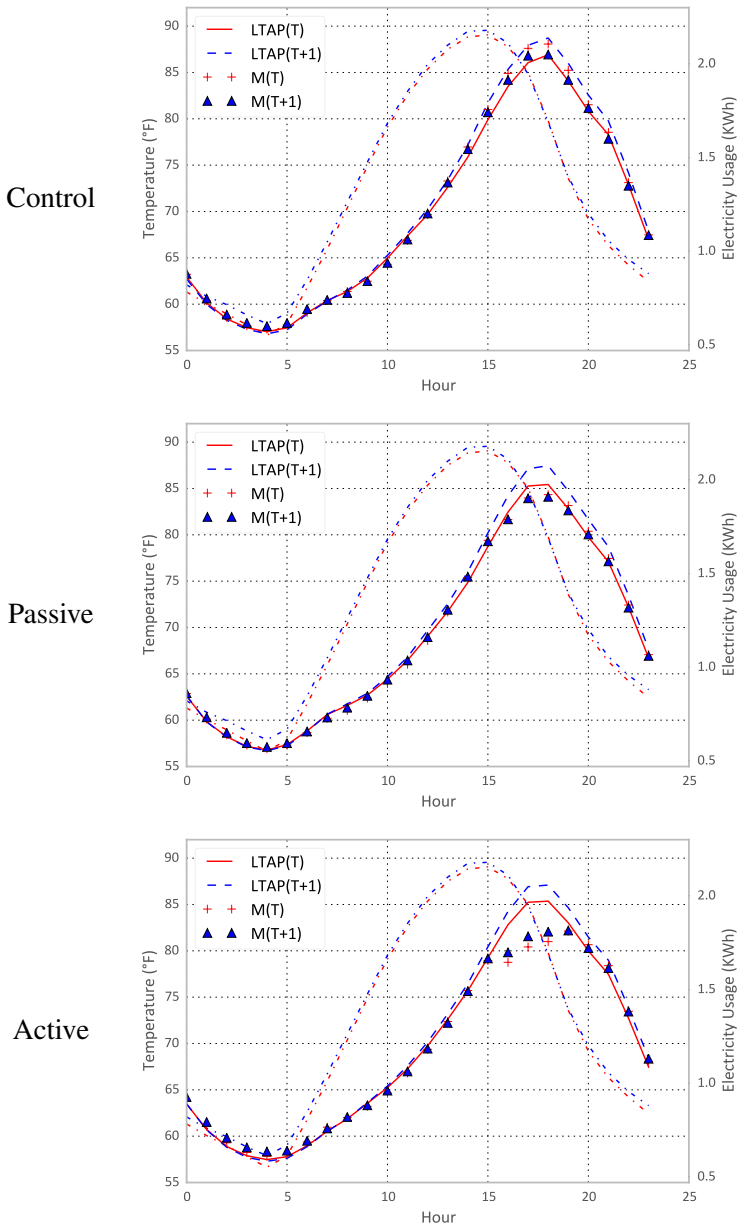

Figure 8: The predicted (by LTAP) and real measured (M) average hourly electricity usage over a day. The predictions have the expected shape and expected delay.

the feature variable for predictions.

Figure 8 shows hourly electricity demand averaged over all summer days in year T and year T+1. In the figure, lines are used to present the predicted values by LTAP, and the individual points are used for actual measured values. Overall, we see that the largest differences appear during the peak-demand hours, where the predicted usage and the real usage are about the same for the control group, while the active groups clearly reduced the usage during the peak-demand hours and the passive group also reduced their usages but not as significantly.

Tables 6 and 7 provide more quantitative measures of the reduction in electricity demand. The LTAP baseline predictions are able to capture the impact of temperature, we can regard the difference between the predicted values and the actual measurements as the "true" measure of energy reduction due to the new pricing schemes. Overall, we see the impact of the new pricing scheme on the overall daily usage is relatively small, while the impact on the usage during peak-demand hours is quite significant.

From Table 7 we see that the active group is able to reduce their usage during the peak-demand hours much more than the passive groups. The reduction by the active groups during the peak-demand hours reaches almost 20%, which is very signficant. There are some households that reduce the usage during peak-demand hours by as much as 40%. This indicates that the new pricing structure is effective in reducing electricity usage during peak-demand hours. It is possible that these active participants choose to opt in because they are better able to respond to the incentives provided by new pricing scheme.

A unexpected observation from this table is that all groups reduced electricity usage in year T+1, even the control group. This particular change in the behavior of the control group appears to explain the decreases in the reduction observed in year T+1 in Table 2. Based on he values in Table 2, we have speculated that the decrease in reduction of electricity usage indicates the active participants have become tired of responding to the changing price during the day. The new baseline with LTAP seems to suggest a new interpretation of the consumer behavior. The control group must have heard about the new behavior of the active participants and started to mimic their behavior even though there is no incentive for them to do so.

7 Summary and Future Work

We set out to study options of derive baseline models from data because the randomized control group is hard to design and is even impossible in some cases. Ultimately, we would like to design a strategy that could generate baseline models for individual participants of a study, while the randomized control group can only serve as the baseline for a whole group. For this work, we have chosen a data set from a well-designed field study of residential electricity usage because it contains a control group that we could compare our baseline model against.

In this work, we explored a number of black-box approaches such as linear regression and Gradient Boosting. Among these machine learning methods, we found Gradient Tree Boosting to be more effective than others. However, the most accurate GTB models are produced with lagged variables as features, for example, the electricity usage a day before and a week before. In order to use the model established on data from year T-1 to make predictions for year T, the existing structure of the prediction procedure effectively requires the actual usage data from year T in order to make predictions for values in year T. We have attempted to modify the prediction procedure to use the recently predictions in place of the actual measured values, however the tests show that the prediction errors accumulated over time, leading to unrealistic predictions a month or so into the summer season. This type of accumulation of prediction errors is common to sequential prediction procedures for time series.

To address the above difficulty, we devised a number of white-box approaches. The method known as LTAP is reported here. It is based on the fact that the aggregated variable electricity usage per day is accurately described by a piece-wise linear function of average daily temperature. This fact allows us to make predictions about the total daily electricity usage. By assuming the usage profile remains the same during the study, we are able to assign the hourly usage values from the aggregated daily usage. This approach is shown to be self-consistent, that is the prediction procedure exactly reproduces the electricity usage in year T-1 and the prediction for the control in year T is very close to the actual measured values. As one might expect, both treatment groups have reduced electricity usage during the peak-demand hours and

the active group reduced the usage more than the passive group.

The analysis results also contain a unexpected revelation, the control group actually reduced its electricity usages in year T+1, the second year after the introduction of the new pricing structures. Previously, using the randomized control group as the baseline, researchers have concluded that there was a decrease in the reduction of the electricity usage during the peak-demand hours. This decrease might be an indication that the new pricing scheme has lost its attractiveness. The new analysis results suggest alternate possibilities, for example, households might have acquired more energy efficient airconditioners, the change the fixed rate at the beginning of year T+1 might have make the consumers more concerned about their electricity usage, or participants of the control group might have adapted the behavior of the treatment groups.

The above hypothesis should be investigated and we are interested in further verify the effectiveness of LTAP. One way to improve LTAP might be to capture additional features, such as the day of the week and so on. So far, we have only considered the average usages of groups, LTAP could be used to make prediction of individual household. We plan to exercise this feature, which might provide additional ways to verify the new baseline model. From our tests on GTB, we noted that the prediction errors seem to accumulate over time, it is of great theoretical interest to study sequential prediction methods that would not accumulate prediction errors over time.

Acknowledgment

The authors gratefully acknowledge the helpful discussions with Sam Borgeson, Daniel Fredman, Liesel Hans, Ling Jin, and Sid Patel.

This work is supported in part by the Director, Office of Laboratory Policy and Infrastructure Management of the U.S. Department of Energy under contract No. DE-AC02-05CH11231. This work is also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT & Future Planning (MSIP) (NRF-2014R1A1A1002662) and the NRF grant funded by the MSIP (NRF-2014M2A8A2074096).

References

Hesham K. Alfares and Mohammad Nazeeruddin. Electric load forecasting: Literature survey and classification of methods. *International Journal of Systems Science*, 33(1):23–34, January 2002.

Peder Bacher and Henrik Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 43(7):1511–1522, 2011.

Vincenzo Bianco, Oronzio Manca, and Sergio Nardini. Electricity consumption forecasting in Italy using linear regression models. *Energy*, 34(9):1413–1421, 2009.

Sam Borgeson. *Targeted Efficiency: Using Customer Meter Data to Improve Efficiency Program Outcomes*. PhD thesis, UC Berkeley, 2014. Available at `http://pqdtopen.proquest.com/pubnum/3686197.html` and `https://escholarship.org/uc/item/32q1w1sf`.

A. Capasso, W. Grattieri, R. Lamedica, and A. Prudenzi. A bottom-up approach to residential load modeling. *IEEE Transactions on Power Systems*, 9(2):957–964, May 1994.

Peter Cappers, Annika Todd, Michael Perry, Bernard Neenan, and Richard Boisvert. Quantifying the impacts of time-based rates, enabling technology, and other treatments in consumer behavior studies: Protocols and guidelines. Technical Report LBNL-6301E, Lawrence Berkeley National Laboratory, 2013. URL `http://eetd.lbl.gov/sites/all/files/lbnl-6301e.pdf`.

Bo-Juen Chen, Ming-Wei Chang, and Chih-Jen Lin. Load forecasting using support vector machines: A study on EUNITE competition 2001. *IEEE Transactions on Power Systems*, 19(4):1821–1830, 2004.

Gianfranco Chicco, Roberto Napoli, and Federico Piglione. Comparisons among clustering techniques for electricity customer classification. *IEEE Transactions on Power Systems*, 21(2):933–940, 2006.

Gianfranco Chicco, Roberto Napoli, Federico Piglione, Petru Postolache, Mircea Scutariu, and Cornel Toader. Load pattern-based classification of electricity customers. *IEEE Transactions on Power Systems*, 19(2):1232–1239, 2004.

John Concato, Nirav Shah, and Ralph I. Horwitz. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, 342(25):1887–1892, June 2000.

William Jay Conover and WJ Conover. *Practical nonparametric statistics*. Wiley New York, 1980.

Dora L. Costa and Matthew E. Kahn. Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, 11(3):680–702, June 2013.

Remy Cottet and Michael Smith. Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849, 2003.

Richard E Edwards, Joshua New, and Lynne E Parker. Predicting future hourly residential electrical consumption: A machine learning case study. *Energy and Buildings*, 49:591–603, 2012.

James A. Espey and Molly Espey. Turning on the lights: A meta-analysis of residential electricity demand elasticities. *Journal of Agricultural and Applied Economics*, 36:65–81, 2004.

Marcelo Espinoza, Caroline Joye, Ronnie Belmans, and Bart De Moor. Short-term load forecasting, profile identification, and customer segmentation: a methodology based on periodic time series. *IEEE Transactions on Power Systems*, 20(3):1622–1630, 2005.

Margaret F Fels. Prism: an introduction. *Energy and Buildings*, 9(1-2):5–18, 1986.

Vera Figueiredo, Fátima Rodrigues, Zita Vale, and Joaquim Borges Gouveia. An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems*, 20 (2):596–602, 2005.

Dipl-Wi-Ing Christoph Flath, Dipl-Wi-Ing David Nicolay, Tobias Conte, PD Dr Clemens van Dinther, and Lilia Filipova-Neumann. Cluster analysis of smart metering data. *Business & Information Systems Engineering*, 4(1):31–39, 2012.

Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Proceedings of International Conference on Machine Learning*, volume 96, pages 148–156, 1996.

Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Ralph I Horwitz and Alvan R Feinstein. Methodologic standards and contradictory results in case-control research. *The American journal of medicine*, 66(4):556–564, 1979.

Samuel Humeau, Tri Kurniawan Wijaya, Matteo Vasirani, and Karl Aberer. Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households. In *Proceedings of Sustainable Internet and ICT for Sustainability*, pages 1–6. IEEE, 2013.

GW Irwin, W Monteith, and WC Beattie. Statistical electricity demand modelling from consumer billing data. In *IEE Proceedings C (Generation, Transmission and Distribution)*, volume 133, pages 328–335. IET, 1986.

Paul L Joskow. California's electricity crisis. *Oxford Review of Economic Policy*, 17(3):365–388, 2001.

Jungsuk Kwac, June Flora, and Ram Rajagopal. Household energy consumption segmentation using hourly data. *IEEE Transactions on Smart Grid*, 5(1):420–430, 2014.

Jungsuk Kwac and Ram Rajagopal. Demand response targeting using big data analytics. In *Proceedings of IEEE International Conference on Big Data*, pages 683–690. IEEE, 2013.

Jeannine Liddle, Margaret Williamson, Les Irwig, and New South Wales. *Method for evaluating research & guideline evidence*. NSW Department of Health, 1996.

F. J. Nogales, J. Contreras, A. J. Conejo, and R. Espinola. Forecasting next-day electricity prices by time series models. *IEEE Transactions on Power Systems*, 17(2):342–348, May 2002.

F. Oldewurtel, A. Ulbig, A. Parisio, G. Andersson, and M. Morari. Reducing peak electricity demand in building climate control using real-time pricing and model predictive control. In *Proceeding of IEEE Conference on Decision and Control*, pages 1927–1932, December 2010.

Anastasios Panagiotelis and Michael Smith. Bayesian density forecasting of intraday electricity prices using multivariate skew t distributions. *International Journal of Forecasting*, 24(4):710–727, 2008. URL `http://www.sciencedirect.com/science/article/pii/S0169207008001015`.

A. Rabl and A. Rialhe. Energy signature models for commercial buildings: test with measured data and interpretation. *Energy and Buildings*, 19(2):143–154, 1992. URL `http://www.sciencedirect.com/science/article/pii/0378778892900085`.

Sarvapali D. Ramchurn, Perukrishnen Vytelingum, Alex Rogers, and Nicholas R. Jennings. Putting the 'smarts' into the smart grid: A grand challenge for artificial intelligence. *Communications of the ACM*, 55(4):86–97, April 2012.

Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

Brian Artur Smith, Jeffrey Wong, and Ram Rajagopal. A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting. In *Proceedings of ACEEE Summer Study on Energy Efficiency in Buildings*, 2012.

L. Suganthi and Anand A. Samuel. Energy models for demand forecasting – a review. *Renewable and Sustainable Energy Reviews*, 16(2):1223–1240, 2012. URL http://www.sciencedirect.com/science/article/pii/S1364032111004242.

Lukas G. Swan and V. Ismet Ugursal. Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, 13(8):1819–1835, 2009.

J. W. Taylor and P. E. McSharry. Short-term load forecasting methods: An evaluation based on European data. *IEEE Transactions on Power Systems*, 22(4):2213–2219, November 2007.

James W. Taylor. Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, 204(1):139–152, 2010. URL http://www.sciencedirect.com/science/article/pii/S037722170900705X.

Annika Todd, Michael Perry, Brian Smith, Michael J. Sullivan, Peter Cappers, and Charles A. Goldman. Insights from smart meters: The potential for peak hour savings from behavior-based programs. Technical Report LBNL-6598E, Lawrence Berkeley National Laboratory, 2014. URL http://escholarship.org/uc/item/2nv5q42n.

Annika Todd, Elizabeth Stuart, Steven R. Schiller, and Charles A. Goldman. Evaluation, measurement, and verification (EM&V) of residential behavior-based energy efficiency programs: Issues and recommendations. Technical Report DOE/EE-0734, US Department of Energy, 2012. URL http://eetd.lbl.gov/sites/all/files/publications/behavior-based-emv.pdf.

George J Tsekouras, Nikos D Hatziargyriou, and Evangelos N Dialynas. Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Transactions on Power Systems*, 22(3):1120–1128, 2007.

Sergio Valero Verdú, Mario Ortiz Garcia, Carolina Senabre, Antonio Gabaldón Marin, and Francisco J García Franco. Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. *IEEE Transactions on Power Systems*, 21(4):1672–1682, 2006.

Frank A Wolak. Diagnosing the California electricity crisis. *The Electricity Journal*, 16(7):11–37, 2003.