Predicting baseline for analysis of electricity pricing

Taehoon Kim, Jaesik Choi, Dongeun Lee, Alex Sim, C. Anna Spurlock, Annika Todd and Kesheng Wu*

Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA
Email: carpedm20@gmail.com
Email: jaesik@unist.ac.kr
Email: eundong@lbl.gov
Email: asim@lbl.gov
Email: caspurlock@lbl.gov
Email: atodd@lbl.gov
Email: kwu@lbl.gov
*Corresponding author

Abstract: To understand the impact of a new pricing structure on residential electricity peak demands, we need a baseline model that captures every factor other than the new price. The gold standard baseline is a randomised control trial, however, control trials are hard to design. The alternative to learn a baseline model from the past measurements fails to make reliable predictions about the daily peak usage values next summer. To overcome these shortcomings, we propose several new methods. Among these methods, the one named LTAP is particularly promising. It accurately predicts future usages of the control group. It also predicts the reductions of the peak demands to remain the same, while previous studies have found the reduction to be diminishing over time. We believe that LTAP is capturing the self-selection bias of the treatment groups better than techniques used in previous studies and are looking for opportunities to confirm this feature.

Keywords: baseline model; residential electricity consumption; outdoor temperature; gradient tree boosting; GTB; electricity rate scheme.


Biographical notes: Taehoon Kim received his BS in Computer Science and Engineering at the Ulsan National Institute of Science and Technology in 2015. He worked as a research intern at the Lawrence Berkeley National Laboratory during the summer of 2015. He is currently working at the Devisisters, a mobile game company, and building a deep reinforcement learning model to automate a game balance testing. His research interests focus on reinforcement learning and deep learning.

Jaesik Choi is an Assistant Professor in the School of Electrical and Computer Engineering at UNIST (Ulsan, Korea) and an affiliate researcher of the Lawrence Berkeley National Laboratory since July 2013. His research focuses on learning and inference with probabilistic graphical models, combining robot motion planning with AI planning, and spatio-temporal data analysis. He received his PhD in Computer Science from the University of Illinois at Urbana-Champaign in 2012 and BS in Computer Engineering from the Seoul National University in 2004.

Dongeun Lee received his PhD in Electrical Engineering and Computer Science and BS in Computer Science and Engineering from the Seoul National University, Korea in 2014 and 2006, respectively. He is a Computer Systems Engineer at the Lawrence Berkeley National Laboratory and also jointly affiliated with the School of Electrical and Computer Engineering at Ulsan National Institute of Science and Technology, as a Postdoctoral Research Associate. His research interests lie at the intersection of statistical signal processing and machine learning. In particular, he focuses on various aspects of big data management and analysis as applications.

Alex Sim is a Senior Computing Engineer at the Lawrence Berkeley National Laboratory. His current R&D activities focus on data mining and modelling, data analysis methods, distributed resource management, and high performance data systems. He authored and co-authored more than 100 technical publications, and released a few software packages under open source license.
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 provides an important proving ground for the predictive analysis capability based on big data analysis platforms (Geerdink, 2015; Ma et al., 2015; Wlodarczyk and Hacker, 2014). These cutting-edge data science techniques together with behavioural theories enable behaviour analytics: novel insights into patterns of electricity consumption and their underlying drivers (Costa and Kahn, 2013; Todd et al., 2014). The baseline models are generated from AMI. Our work focuses on extracting baseline electricity usage predictions based on an observation about daily peak usage. The process of designing these pricing schemes, recruiting participants for a pilot study, implementing the pricing schemes, and monitoring the impacts takes a few years. The baseline model needs to capture consumer behaviour prior to the new pricing schemes, and then predict what consumer behaviour would be without the pricing changes. This is challenging because the baseline model not only needs to predict intraday electricity usage, but also needs to be accurate for years into the future. Furthermore, in preliminary tests, we have noticed the impact of the pricing schemes to be weaker than the impact of other factors such as temperature. Therefore, the baseline model must incorporate the outdoor temperature, which has a complex relationship with the electricity demand.

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 behaviour that persists for a long time, while the forecast model typically aims to forecast for the next few cycles of the time series. Usually, techniques that make forecasts for years into the future are based on highly aggregated time series with months 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 motivated our work, the overall objective is to study the impacts of pricing policies on daily peak usage. The process of designing these pricing schemes, recruiting participants for a pilot study, implementing the pricing schemes, and monitoring the impacts takes a few years. The baseline model needs to capture consumer behaviour prior to the new pricing schemes, and then predict what consumer behaviour would be without the pricing changes. This is challenging because the baseline model not only needs to predict intraday electricity usage, but also needs to be accurate for years into the future. Furthermore, in preliminary tests, we have noticed the impact of the pricing schemes to be weaker than the impact of other factors such as temperature. Therefore, the baseline model must incorporate the outdoor temperature, which has a complex relationship with the electricity demand.

This work develops a number of new baseline models that could satisfy the above requirements. In the class of black-box methods, we extend the most promising regression method from the previous study (Kim et al., 2015) to make continuous predictions. In the class of white-box methods, we develop a technique to make hourly usage predictions based on an observation about daily

---

C. Anna Spurlock received her PhD in Agricultural and Resource Economics from the University of California at Berkeley in 2013. She currently works at the Lawrence Berkeley National Laboratory conducting research on a variety of topics including energy efficiency policy and consumer behaviour.

Annika Todd is a Senior Scientific Engineering Associate in the Electricity Markets and Policy Group at Lawrence Berkeley National Laboratory. She is an experimental and behavioural economist, and conducts research and analysis on energy efficiency, demand response, and smart grid topics including smart grid investment grant (SGIG) dynamic pricing projects, evaluation, measurement and verification of energy efficiency programs, and technical assistance to states on energy efficiency programs.

Kesheng Wu is a Senior Computer Scientist at Lawrence Berkeley National Laboratory. He works actively on a number of topics in data analysis, data management, and high-performance computing. He has authored and co-authored more than 100 technical publications.

This paper is an extended and revised version of the paper ‘Extracting baseline electricity usage with gradient tree boosting’ presented at DataCom 2015, Chengdu, China, 19–21 December 2015.
aggregate electricity usage and temperature. At present, the gold standard baseline is a well-designed randomised control group. We aim to show that our new data-driven baselines could accurately predict the average electricity usage of the control group. For this evaluation, we use a well-designed study from a region of the USA where the electricity usage is the highest in the afternoon and evening during the months of May through August. Though this work concentrates on showing the new baseline models are effective for groups, we believe that these new models are also useful for studying individual households in the future.

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 of 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 behaviour such as electricity consumption segmentation (Chicco et al., 2004, 2006; Figueiredo et al., 2005; Verdú 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 maximise 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 behaviour. 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, because increased outdoor temperature leads to increased indoor temperature, which in turn leads people to turn on their air-conditioners, 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 effects on electricity usage, and others, such as number of occupants in the building, clearly affect electricity usage but their values are unknown or their impact on electricity usage is multifaceted or unknown (Borgeson, 2014; Fels, 1986; Rabl and Rialhe, 1992). For this reason, many researchers refer to these models as ‘grey-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 behaviour 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 aggregate demand and therefore could parameterise 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 use patterns 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 randomised control trial (Cappers et al., 2013; Todd et al., 2014). 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 the electricity usage (from June 1 to August 31), yielding 41,698,080 data records. The data records from the 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 randomised control trial, the control group is a randomly selected set of households that are meant to be used as the baseline (Costa and Kahn, 2013; Concato et al., 2000). This control group is unaware of the study and stays with the previously available fixed-price scheme throughout the testing period.1

There were a number of treatment groups based on the pricing structures and other factors. To simplify the discussion in this paper, we concentrate on two groups following one particular time-based price, where during the peak-usage hours, 4 PM to 7 PM in the region of this study, the per-kWh charge is higher than the rest of the day. In the active group, households are invited to participate in the pilot study involving the new pricing structure, and have to opt in to the study by responding positively to the invitation. While the households in the passive group are informed of their participation in the new pricing trial, and offered a chance to opt out of the trial.

The designers of the study have taken account of the expected response rates of the two groups and sent out a lot more invitations for households to opt in than the invitations for households to opt out. Even with this design, the number of active participants (those who explicitly opt in) and the number of passive participants (those who did not opt out) are quite different. To avoid the imbalance between the two groups, we randomly selected about 4,000 households from each. Over the three years of study, about 26% of households moved or otherwise changed, and we dropped these households that do not have measurement data for the whole duration of the study.

3.2 Overall statistics
Figure 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 Figure 2, where average hourly temperatures and electricity usages are plotted against hour.

In Figure 2, the temperatures of year \(T\) and \(T+1\) are higher than the temperature of year \(T-1\), and the electricity usage increases in year \(T\) and \(T+1\) are higher as well. To see the impact of the new pricing scheme, which was designed to reduce electricity usage in the peak hours, we need to carefully capture the effect of temperature. From Figure 2, we observe the impact of temperature on electricity usage is not instantaneous, rather it is effect appears to be delayed for a few hours. The increased electricity usage during the summer afternoon is mostly from air-conditioning, which is more directly related to the indoor temperature, while the temperature reported in our dataset is the outdoor temperature. It takes time for the increased outdoor temperature to raise the indoor temperature.

Because there are no obvious differences from Figure 1 and Figure 2, we conclude that the influence of common features such as season, outdoor temperature, day of the week and so on are stronger than the features that distinguish the groups. This means the baseline models have to be very accurate in order to isolate the effect of the new prices between the treatment groups and the control group. We will discuss these methods carefully in Section 4.

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Average hourly usage</th>
<th>Subtract control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(T-1)</td>
<td>(T)</td>
</tr>
<tr>
<td>Control</td>
<td>1.128</td>
<td>1.205</td>
</tr>
<tr>
<td>Passive</td>
<td>1.100</td>
<td>1.152</td>
</tr>
<tr>
<td>Active</td>
<td>1.125</td>
<td>1.160</td>
</tr>
</tbody>
</table>

3.3 Comparison against the control group
In the tradition of randomised 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 randomised 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 randomised control group. We are interested in developing such a methodology for two reasons:

1 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 counterfactual control group

2 it is often the case that programs, such as the pricing programs used in this paper, are implemented by electricity providers without a randomised evaluation methodology.
Figure 1  Daily electricity usages of three groups for year \((T - 1, T, T + 1)\) (see online version for colours)
In the rest of this section, we demonstrate the problems that would emerge if a utility were to attempt to evaluate a new rate or program by comparing only those households that selected into the program to those households that did not.

We do this by making the comparison between the control group and the two self-selected treatment groups. 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 are averaged over all hours and all days of the summer months in each year, while the values in Table 2 are 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.

To further illustrate the effect of self-selection bias, we next examine whether the electricity usage 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 groups has about 4,000 households, therefore, we expect the confidence interval of the these average values to be about $0.85 / \sqrt{4,000} = 0.013$. If each of the three groups contain randomly selected households, the differences between these groups should be less than 0.013. From Table 1, we see that the absolute difference between the active group and the control group is smaller than 0.013, while the absolute difference between the passive group and the control group is larger than 0.013. This might suggest that the passive group is different from the control group. Next, we examine the time series to see whether there is any large differences among them.

Table 3 shows the Kolmogorov-Smirnov test scores from comparing the hourly electricity time series over three summers. The two time series are considered as likely to be generated from the same probability density distribution.

### Table 3

<table>
<thead>
<tr>
<th>Year</th>
<th>$T-1$</th>
<th>$T$</th>
<th>$T+1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control v. Passive</td>
<td>0.09</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Control v. Active</td>
<td>0.01</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Passive v. Active</td>
<td>0.09</td>
<td>0.02</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Note: 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.

#### 3.4 Comparing the time series

Next, we directly compare the time series of the average hourly usage of each group to better understand their differences. For this test, we have selected the Kolmogorov-Smirnov test (KS test) (Conover and Conover, 1980). Given two time series, the KS test measures the distance between their 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 received the same pricing scheme, we expect the control group to behave similar to the treatment groups. In terms of KS test scores, we anticipate all three KS test scores to be greater than 0.05. However, only two of the three scores are greater than 0.05. The KS test between the control group and the passive group produced a score of 0.09, which is clearly larger than 0.05, indicating these two time series should be considered the ‘same’. Even though the difference of their average daily usage values appear to be somewhat large in Table 1, we believe the KS test provide a more holistic comparison of the two time series and regard the passive group to be statistically the same as the control group. This provides a strong evidence that the control group could be used as the baseline for the active group.

The KS test between the active group and the control group produced a score of 0.01, much less than 0.05, which indicates that these two groups are different. Based on the same belief that the KS test provides a more holistic measure of the two sets of electricity usage values, we tend to believe that active group and the control group are different although their average daily usages are nearly the same as shown in Table 1. This difference casts doubt on the suitability of using the control group as the baseline for the active group. This doubt is the key motivation for us to develop data-driven baselines.

The KS test between the passive group and the active group produced a score of 0.09, indicating the two time series are the ‘same’, and their average values are also fairly close to each other as shown in Table 1.

The KS test scores for year $T$ and year $T + 1$ in Table 3 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 behaviour differences.

## 4 Methodology

The statistics provided in the previous section suggest that the treatment groups in this study have strong self-selection bias, therefore, directly comparing the randomised control group with the treatment groups might not accurately reflect the true impact of the new pricing structure. This is the key 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 randomised control group is only a good baseline for the average behaviour of a treatment group, not any individual household. In this section, we first introduce a few black-box approaches and then introduce a white-box approach. The black-box methods include a number of statistical machine learning methods, such as linear regression (LR), gradient linear boosting (GLB), and gradient tree boosting (GTB). The white-box method is built upon a relationship between the temperature and total (aggregate) daily electricity usage. Next, we provide a brief description of each of these methods.

### Figure 3 An example of GTB model (see online version for colours)

Notes: 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.

#### 4.1 Linear regression

One of the popular yet simple regression models is the 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 dataset \( \{y_i, x_{i,1}, \ldots, x_{i,K}\} \) of \( n \) statistical units, an LR can be represented as follows:

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

where \( \hat{y}_i \) is an estimated value of \( y_i \), \( \beta_k \) is the \( k \)th regression coefficient of \( x_{i,k} \), and \( \epsilon \) is an error term.

#### 4.2 GLB and GTB

Boosting is a prediction algorithm that combines 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 (Schapire, 1990; Freund et al., 1996; Schapire and Freund, 2012).

Gradient boosting (GB) is one of the boosting methods that construct an additive regression model by sequentially training weak learners in the gradient descent viewpoint (Friedman, 2001). GB can be further distinguished by choosing different weak learners. Here, we choose two different weak learners: linear function and decision tree. Each model is called GLB and GTB, respectively. Figure 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:

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

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 minimise the following objective function \( \text{obj}(\Theta) \) of \( \Theta \) which denotes the parameters of GB:

\[
\text{obj}(\Theta) = L(\Theta) + \sum_{k=1}^{K} \Omega(f_k),
\]

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

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

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

### 4.3 Linear relation between temperature and aggregate power (LTAP)

Next, we describe an effective white-box model in our tests. It is well-known that 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 aggregate electricity usage and the average outdoor 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 the outdoor temperature and the hourly electricity usage is complex, but the daily average temperature and the average outdoor temperature is relatively straightforward. Since this work is primarily concerned with the peak usage during the summer, the demands of air-conditioners dominate the electricity usage. From the earlier studies on 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 seasons. The usage that is beyond the minimum varies from hour to hour; we call this portion the variable electricity usage. For the region where these data were recorded, we assume the primary demand for this variable usage is from air-conditioners, and is therefore related to the outdoor temperature.

The reason that the aggregate variable electricity usage is likely to be 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 air-conditioner 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, more heat would 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 relationship with the average outdoor temperature. From the research literature and our own tests presented in the next section, we see that this is true (Rabl and Rialhe, 1992; Bacher and Madsen, 2011; Borgeson, 2014). 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 aggregate 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 \( t_i \) of the day from the hourly 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 values 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 \( b_0 \). We compute \( a_1 \) as follows:

\[
a_1 = a_0 + s (t_i - t_0).
\]

We assign the hourly electricity usage as follows:

\[
h_i[i] + b_0 + (h_0[i] - b_0) a_1 / a_0.
\]

It is easy to verify that the above assignment of the aggregate 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 without carrying the prediction errors from earlier days in the time series.

### 5 Black-box regression models

To establish a baseline through machine learning techniques, we need to first determine the features that this model depends on. From information in the literature and our exploration of the dataset (Rabl and Rialhe, 1992; Bacher and Madsen, 2011; Borgeson, 2014), we choose eight features: three time variables (month, hour, and day of week), two historical electricity usage variable (electricity usage of the same hours on a day before (yesterday) and a week before), and three hourly weather conditions (temperature, atmospheric pressure, and dew point). The role of the historical usage data is to distinguish one
Predicting baseline for analysis of electricity pricing

The weather conditions such as temperature, atmospheric pressure and the dew point could significantly affect the uses of air-conditioners and would also vary in a large geographical area. Unfortunately, the sample data contains only a single set of measurements of weather conditions. This lack of more detailed data on weather related variables makes it hard to fully explore the impact of these variables on electricity usage with this dataset. There are many other features that could affect the electricity usage of a household, such as whether or not the house has an air-conditioner, however, this information is either missing completely or only available for a small number of households. Due to these reasons, our black-box models will only use the above mentioned eight features.

### 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 behaviour. 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 the tree surpassed 5. For each step, we randomly collected half of the dataset 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 five-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 and GLB are larger than those of GTB. This is not unexpected since the relationship between the electricity usage and the temperature is not only nonlinear but also delayed. In the rest of this paper, we choose GTB as the representative of the black-box baseline models.

### 5.2 Training GTB

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 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.

Figure 4 shows the f-score of each feature in GTB, where the f-score is the number of appearances of a feature in all of the 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 the week-before usage) and the third most influential feature is temperature. In Figure 4, we can see how GTB finds which features are important. It is interesting to note that ‘day of week’ is not as effective as other features. This surprised us because we originally assumed that GTB might detect the difference between weekend and weekday from the dataset.

**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 (see online version for colours)
Figure 5  Predicted (by GTB) and measured hourly average electricity usage during year $T$ and $T + 1$ (see online version for colours)

Notes: The lines and symbols representing data from the same year have the same colour. The measured values are lower than the predictions indicating the consumers have reduced electricity usages compared to the ‘business-as-usual’ predictions.
5.3 Hourly averaged prediction

Figure 5 shows the hourly usage prediction by GTB and hourly average temperature. In year $T$ and $T+1$, we see that the control group uses just about the same amount of electricity as predicted by the GTB model, while the treatment groups use less electricity, especially during the peak-demand hours, than predicted by the GTB models. Furthermore, we see that the points representing the measured usages are noticeably below the lines representing the predictions. Clearly, the new pricing scheme has an impact on consumer behaviour, and the active group responded more than the passive group. This is consistent with other research that has examined the response to this type of program differentiated by active or passive enrolment approaches (Cappers et al., 2016). We also see that the GTB model effectively has learned the lagged effect of temperature explained in Figure 2.

Note: 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.
5.4 Continuous predictions with GTB

The features used for our GTB model include the electricity usage from a day ago and a week ago, while we make predictions months or years into the future. 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 as known values, we should be able to supply the values of these lagged variables as well. When making predictions for June of year $T$, we can assume to know all the values before June 1st. When making predictions for June 1st, we can make use of values from May. However, when making predictions for June 2nd, the values from ‘yesterday’ (i.e., June 1st) are not supposed to be available. However, the existing GTB implementation requires the values for all ‘yesterday’s’ to be available before any prediction is made. Clearly, we need to modify GTB to avoid referring to values that are not available yet. The basic strategy is to make predictions in a progressive manner. When making predictions for June 2nd, we use the predictions for June 1st as the values for ‘yesterday’. Similarly, as time progresses, the predicted values will be used for ‘week before’. Effectively, we have produced a continuous version of GTB for making predictions involving lagged variables. This continuous prediction procedure makes predictions one hour at a time and immediately makes use of the predicted values for lagged variables.

Figure 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 increase. This appears to be an accumulation of the 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. Next, we examine a white-box prediction method that successfully avoids this accumulation of prediction errors.

6 White-box prediction

In Section 4.3, we describe the white-box prediction called LTAP. In this section, we first illustrate the linear relationship between the aggregate variable electricity usage and the average daily temperature, which is the basis of the LTAP prediction, and then describe the results of predictions with LTAP. The second half of this section contains the performance results of LTAP.

6.1 Linear relationship between aggregate power usage and temperature

In Section 4.3, we provide some arguments for a linear relationship between the aggregate 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 aggregate variable electricity usage against the average daily temperature. These scatter plots suggest that below 65°F, there is no obvious relationship between 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 air-conditioning is heavily used, it is unsurprising that we observe a simpler relation between temperature and electricity usage.

<table>
<thead>
<tr>
<th>Table 5</th>
<th>The slopes and the coefficients of correlation for data points with average temperature above 65°F from the summer of year $T - 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Slope 1.13</td>
</tr>
<tr>
<td>Passive</td>
<td>1.07</td>
</tr>
<tr>
<td>Active</td>
<td>1.02</td>
</tr>
</tbody>
</table>

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 establish that the relationship between the aggregate electricity usage and the average temperature is 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 influential, such as the day of the week, have been found to be less important. At this time, we only use the temperature as the feature variable for predictions.
Figure 7  Scatter plots of aggregate variable electricity usage and average daily temperature from the summer of year $T-1$ (see online version for colours)
Figure 8  The predicted (by LTAP) and real measured (M) average hourly electricity usage over a day (see online version for colours)

Note: The predictions have the expected shape and expected delay.
control group, as was done in Table 2. The fact that the estimates were consistent with LTAP potentially addressing the passive group are even larger. This observation is because they are better able to respond to the incentives provided by the new pricing scheme.

Tables 6 and 7 provide more quantitative measures of the reduction in electricity demand. Earlier, we showed that the expected variance from random sampling for daily average electricity usage is about 0.013, and from Table 6, the differences between predicted values and the actual measured values are within this variance for the control group. Therefore, we say that the LTAP predictions agree with the actual measurements. Overall, we see the impact of the new pricing scheme on total 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 group. The reduction by the active group during the peak-demand hours reaches almost 10%, which is significant. 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 the new pricing scheme.

The values in Table 7 are in many ways similar to those in Table 2, however, there are some notable differences reflecting the differences between the LTAP baseline model and the randomised control group. The most obvious trend from Table 2 is that the amount of electricity saved during the peak-demand hours is less in year $T + 1$ than in year $T$. This is true for both the active group and the passive group. However, from Table 7, we see the magnitude of usage reduction increased for the passive group and remained the same for the active group. Furthermore, the reductions by the active group in Table 2 are 0.277 kWh for year $T$ and 0.198 kWh for year $T + 1$, which are 60% and 21% larger than 0.164 reported in Table 7. The relative differences for the passive group are even larger. This observation is consistent with LTAP potentially addressing the self-selection bias introduced by estimating reductions by simply comparing the self-selected treated customers to the control group, as was done in Table 2. The fact that the estimated reductions, when calculated using the LTAP baseline, are smaller, suggests that self-selection biased the reductions away from zero in the Table 2 results, which is what we would expect. However, additional studies are needed to confirm this conjecture.

<table>
<thead>
<tr>
<th>Group</th>
<th>$P_T$</th>
<th>$P_{T+1}$</th>
<th>$M_T - P_T$</th>
<th>$M_{T+1} - P_{T+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1.183</td>
<td>1.199</td>
<td>0.022</td>
<td>-0.002</td>
</tr>
<tr>
<td>Passive</td>
<td>1.153</td>
<td>1.171</td>
<td>-0.002</td>
<td>-0.017</td>
</tr>
<tr>
<td>Active</td>
<td>1.179</td>
<td>1.193</td>
<td>-0.019</td>
<td>-0.020</td>
</tr>
</tbody>
</table>

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 group clearly reduced the usage during the peak-demand hours, and the passive group also reduced their usages but not as much.

Tables 6 and 7 provide more quantitative measures of the reduction in electricity demand. Earlier, we showed that the expected variance from random sampling for daily average electricity usage is about 0.013, and from Table 6, we see the differences between predicted values and the actual measured values are within this variance for the control group. Therefore, we say that the LTAP predictions agree with the actual measurements. Overall, we see the impact of the new pricing scheme on total 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 group. The reduction by the active group during the peak-demand hours reaches almost 10%, which is significant. 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 the new pricing scheme.

The values in Table 7 are in many ways similar to those in Table 2, however, there are some notable differences reflecting the differences between the LTAP baseline model and the randomised control group. The most obvious trend from Table 2 is that the amount of electricity saved during the peak-demand hours is less in year $T + 1$ than in year $T$. This is true for both the active group and the passive group. However, from Table 7, we see the magnitude of usage reduction increased for the passive group and remained the same for the active group. Furthermore, the reductions by the active group in Table 2 are 0.277 kWh for year $T$ and 0.198 kWh for year $T + 1$, which are 60% and 21% larger than 0.164 reported in Table 7. The relative differences for the passive group are even larger. This observation is consistent with LTAP potentially addressing the self-selection bias introduced by estimating reductions by simply comparing the self-selected treated customers to the control group, as was done in Table 2. The fact that the estimated reductions, when calculated using the LTAP baseline, are smaller, suggests that self-selection biased the reductions away from zero in the Table 2 results, which is what we would expect. However, additional studies are needed to confirm this conjecture.

Table 7 The average hourly electricity demand during peak-demand hours and their differences from the actual measurements

<table>
<thead>
<tr>
<th>Group</th>
<th>$P_T$</th>
<th>$P_{T+1}$</th>
<th>$M_T - P_T$</th>
<th>$M_{T+1} - P_{T+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1.904</td>
<td>1.957</td>
<td>0.069</td>
<td>-0.020</td>
</tr>
<tr>
<td>Passive</td>
<td>1.849</td>
<td>1.897</td>
<td>-0.027</td>
<td>-0.080</td>
</tr>
<tr>
<td>Active</td>
<td>1.860</td>
<td>1.903</td>
<td>-0.164</td>
<td>-0.164</td>
</tr>
</tbody>
</table>

Note: The predictions are made with LTAP.

7 Summary and future work
We set out to study options for deriving baseline models from data because a randomised control group is not always easy or feasible to obtain. In addition, we would like to design a strategy that could generate baseline models for individual participants in a study, while the randomised control group can only serve as the baseline for the whole group. For this work, we have chosen a dataset from a well-designed field study of residential electricity usage because it contains a control group that we could compare our baseline model against.

We explored a number of black-box approaches such as LR and GB. Among these machine learning methods, we found GTB to be more effective than others. However, the most accurate GTB models require 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 requires the actual usage data from year $T$ in order to make predictions for values in year $T$. We attempted to modify the prediction procedure to use the recent predictions in place of the actual measured values, however our tests show that the prediction errors accumulate over time, leading to unrealistic predictions a month or so into the summer season. This type of accumulation of prediction errors is common to continuous prediction procedures for time series.

To address the above difficulty, we devised a number of white-box approaches, the most effective of which, known as LTAP, is reported here. LTAP is based on the fact that the aggregate 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 further assuming the usage profile of each household remains the same during the study, we are able to assign the hourly usage values from the daily aggregate 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 predictions for the control group in both year $T$ and $T + 1$
are very close to the actual measured values. Both treatment groups have reduced electricity usages during the peak-demand hours and the active group reduced the usage more than the passive group. This observation is consistent with other studies (Cappers et al., 2016).

Though the new data-driven baseline model LTAP predicts the average usages of the control group accurately, there are some differences in predicted impact of the new time-of-use pricing intended to reduce the usage during the peak-demand hours. For example, compared to the control group, the active group reduce its usage by 0.277 kWh (out of about 2 kWh) averaged over the peak-demand hours in the first year with the new price and 0.198 kWh in the second year. Using the LTAP baseline, the average reductions are only 0.164 kWh for both years. Part of the difference may be due to the self-selection bias in treatment groups, especially the active group, where the households have to explicitly opt in to participate in the trial. It is likely that the households that elected to join the active group are well-suited to take advantage of the proposed new pricing structure. We believe that the LTAP baseline is a way to address the self-selection bias, and plan to conduct additional studies to further verify this capability.

Ultimately, we would like to develop a baseline model for each household. Significant work is needed to quantify the suitability of LTAP for this purpose.

Acknowledgments

The authors gratefully acknowledge the helpful discussion 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 US 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 and Future Planning (MSIP) (NRF-2014R1A1A1002662) and the NRF grant funded by the MSIP (NRF-2014M2A8A2074096).

References


Notes
1 There was an adjustment 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 a decrease in the rate. This price change might encourage households to consume more electricity since the incremental cost has decreased.

2 The actual values are 0.83 for Year $T-1$, 0.85 for year $T$, and 0.91 for year $T+1$.

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