Clustering analysis of residential electricity demand profiles

Joshua D. Rhodes^{a,*}, Wesley J. Cole^b, Charles R. Upshaw^c, Thomas F. Edgar^{b,d}, Michael E. Webber^{c,d,**}

^a*Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, 301 E Dean Keeton St. Stop C1752, Austin, TX 78712-1094*

^b*McKetta Department of Chemical Engineering, The University of Texas at Austin, 200 E. Dean Keeton St. Stop C0400, Austin, TX 78712-1589*

^c*Department of Mechanical Engineering, The University of Texas at Austin, 204 E. Dean Keeton Street, Stop C2200, Austin, Texas 78712-1591*

^d*Energy Institute – The University of Texas at Austin, 2304 Whitis Ave Stop C2400 Austin TX 78712-1718*

Abstract

Little is known about variations in electricity use at finely-resolved timescales, or the drivers for those variations. Using measured electricity use data from 103 homes in Austin, TX, this analysis sought to 1) determine the shape of seasonally-resolved residential demand profiles, 2) determine the optimal number of normalized representative residential electricity use profiles within each season, and 3) draw correlations to the different profiles based on survey data from the occupants of the 103 homes. Within each season, homes with similar hourly electricity use patterns were clustered into groups using the *k-means* clustering algorithm. Then probit regression was performed to

*Corresponding author

**Principal corresponding author

Email addresses: joshdr@utexas.edu (Joshua D. Rhodes), wcole@che.utexas.edu (Wesley J. Cole), crupshaw@utexas.edu (Charles R. Upshaw), edgar@mail.utexas.edu (Thomas F. Edgar), webber@mail.utexas.edu (Michael E. Webber)

Preprint submitted to Applied Energy

July 2, 2014

determine if homeowner survey responses could serve as predictors for the clustering results. This analysis found that Austin homes fall into one of two seasonal groups with some homes using more expensive electricity (from a wholesale electricity market perspective) than others. Regression results indicate that variables such as if someone works from home, hours of television watched per week, and education levels have significant correlations with average profile shape, but might vary across seasons. The results herein also indicate that policies such as time-of-use or real-time electricity structures might be more likely to affect lower income households during some high electricity use parts of the year.

Keywords:

energy, smart meter data, residential

1. Introduction

Electricity generation is responsible for 40% of primary energy use in the US [1], and while significant information is available about its generation, much less is known about its end use, particularly on an inter-daily temporal scale. Current smart meter and smart grid deployments are vastly increasing the amount of energy use information being created, curated, and analyzed. These datasets, which typically have 15-minute or 1-hour granularity, increase the amount of available energy use information by orders of magnitude over once-per-month meter reads and enable analyses not previously possible. In particular, they allow for temporal assessment of electricity use, which is important because it holds the potential to reveal non-obvious insights about electricity consumption and the behavioral and technological

drivers of that consumption. Also, the temporal aspect of electricity use is significant because electricity is expensive to store and thus is typically produced at the rate of consumption. This balance that must occur in the wholesale market can cause the price for electricity to be volatile, especially during peak demand times.

Residential power demand can range in magnitudes from a few hundred watts (W) to the low tens of kilowatts (kW) per household, especially in regions where residential air-conditioning units are ubiquitous. Inter-season differences for residential total grid power demand in cooling climates can change by over 400% from non-cooling to cooling periods [2]. Residential electrical use patterns fluctuate differently during the day due to space-conditioning set points, time of year, weather, occupant behavior and schedules. In a similar way, the electric grid load and net carbon emissions change temporally due to changes in demand, temperatures, and generation mixes, which are also dynamic with daily and seasonal timescales [3, 4].

The drivers of macro-level, aggregate energy use are current topics of study [5], but the factors affecting residential energy use at finely-resolved timescales are poorly understood. This work seeks to fill part of that knowledge gap by identifying correlations between electric customer survey data and electricity use profiles. Understanding electricity consumption patterns and their influencing factors could potentially lead to more robust energy modeling, more precise demand forecasts, and more effective energy conservation and peak reduction campaigns. While beyond the scope of this paper, these results might also inform analysis of behavior-related tradeoff calculations such as the environmental impacts of telecommuting [6].

Clustering techniques are used to find distinctive groups of profiles within seasons. Binary regression analysis is then used to draw correlations between homeowner surveys and clustering results. Clustering is an effective tool for analyzing static data or time series data, such as electricity usage [7, 8]. For example, Chicco [9] used a variety of clustering techniques to analyze 400 load patterns of non-residential, medium voltage consumers during a weekday in Italy. The electricity data were in 15-minute intervals and were normalized by the minimum and maximum value of the daily load pattern so that all normalized load profiles were between 0 and 1. He found that choosing the best clustering algorithms depended on whether or not the purpose was to identify the outliers or to assign all the load profiles to a specific category.

Kim *et al.* [10] performed a similar analysis by clustering electricity usage from high voltage consumers in South Korea to determine typical load profiles using 15-minute data. They had 3,183 high voltage customers from the Korea Electric Power Corporation (KEPCO) whose meter data were used for this study. They found the hierarchical clustering technique to be the most effective in determining typical load profiles.

Kwac *et al.* [11] formed a ‘dictionary’ set of hourly electricity use profiles from 220,000 homes in California, for a total of 66 million individual (house-day) profiles, each normalized by its total daily consumption. They used an adaptive k-means clustering algorithm that increased the number of clusters as needed so that a mean squared error threshold was met. They also found that although two homes might have the same average profiles, the ‘entropy’ or difference in raw daily profiles could vary significantly. Given their dataset they were also able to make distinctions between homes in different California

climate zones.

Räsänen *et al.* [12] performed clustering of hourly electricity usage over a more diverse consumer base in Finland. The study included 3,989 consumers in the following categories: 80% residential, 8% public sector, 6% services, 5% agriculture, and 1% industry. To perform the clustering, they reduced the size of the dataset, so they performed the clustering based on 489 randomly chosen time points throughout the year (489 represents 5% of the year). Their data-driven clustering technique provided more accurate load forecasts than the current methods used by the utility.

Zhou *et al.* [13] and Panapakidis *et al.* [14] used smaller datasets for their clustering analyses of electricity loads. Zhou *et al.* used 72 load profiles from six different consumers in China and Panapakidis *et al.* used a mix of 150 residential, commercial, and industrial consumers in Greece. Panapakidis *et al.* notes that a consumer’s electricity costs might also be helpful to consider when performing clustering.

Clustering analysis of temporally-resolved electricity use has currently been limited to generating typical load profiles, grouping typical customers, or finding outliers. This analysis seeks to take the analysis a step further by leveraging the comprehensive data set available from a local smart grid demonstration project [15] and explore the drivers behind the derived groups of electricity customers.

2. Methods

This analysis employed a data-driven approach to 1) create average seasonal curves for each home in the analysis, 2) determine the number of rep-

representative residential electricity demand profiles within their respective seasons, and 3) attempt to draw correlations to the different profiles based on data from the same homes. Initial data averaging, grouping, and home curve derivations were performed using the Texas Advanced Computer Center’s data applications resource Corral [16]. Corral is a collection of storage and data management resources available to researchers at the University of Texas at Austin. Data analytics and clustering were performed using the statistical program *R* (version 3.0.2). Figure 1 provides a flowchart of the analysis progression.

Figure 1: This flowchart shows the datasets and sequence of steps for the analysis.

2.1. Creating average seasonal curves.

The data used in this analysis are collected as part of a smart grid demonstration project located in Austin, TX and operated by the Pecan Street Research Institute [17]. Pecan Street’s Smart Grid Demonstration Project

consists of over 1000 homes that are monitored at the sub circuit (16 circuits per home or better) at a 1-min frequency (or better). Many of these homes also have solar PV installations, EVs, batteries, and smart thermostats. For a full description of the project, see Rhodes et al., 2014 [15]. The subset of data analyzed was power consumption in watts measured every 1 minute for 103 homes in Austin from November 2012 to October 2013. Note that many of the homes in the study have solar PV installed, but the data used here are the home consumption data before solar PV generation is added to the curve. The 1 minute data were used to create a representative electricity use profile for each home for every season of the year. The data were first averaged over each hour to create hourly profiles. For example, electricity use data from 14:01 – 15:00 were averaged and assigned to the value 15:00 (interval ending). This yielded a 24-element vector for each day of the year. Figure 2 shows a 10 day sample of 1 minute data for a single home. The left plot shows the raw 1 minute data and the right plot shows the same data averaged to 1 hour profiles, with the average of the hourly profiles in bold.

The bold curve in Figure 2 is for the month of May for a graphical illustration of how the curves were made. For the rest of this analysis, the curves are seasonal curves based on the four seasons defined as; summer: June – August, fall: September – November, winter: December – February, and spring: March – May given by Equation 1:

Figure 2: Figure showing 10 days (May 01, 2013 – May 10, 2013) worth of the initial, raw 1 minute data from a single home on the left, averaged 1 hour profiles (same days) on the right with the bold line showing the average of the hourly curves for the entire month of May 2013. The raw data (left panel) show the great variability in the demand profiles, whereas the hourly averages and averaged profile (right panel) show the general diurnal variation for consumption.

$$curve_j = \begin{pmatrix} w_{j0} \\ w_{j2} \\ \vdots \\ w_{jh} \\ \vdots \\ w_{j23} \end{pmatrix}, \quad (1)$$

where $curve_j$ is the representative seasonal curve for home j and w_{jh} is the hourly energy consumption of home j at hour h given by Equation 2:

$$w_{jh} = \sum_{m_h} P_{jm}/60 \quad (2)$$

where P_{jm} is the measured average power draw for home j over minute m of hour h .

The primary focus of this research was temporal variation in electricity use (i.e., profile shape), rather than magnitude, which has been considered in other analyses [15]. Therefore each raw, hourly usage curve was normalized by its largest value so that all home profiles would be on a fractional 0 to 1 scale. The minimum value was not subtracted from the curve so as not to exacerbate the difference between minimum and maximum electricity use. For example, some homes, on average reduced their usage to almost zero during certain parts of the day while others never dropped their average usage below 40% of their average peak usage. The normalized seasonal curves are then used as inputs to the optimal clustering algorithm discussed in the next section.

2.2. Determining distinctive clusters within seasons.

After creating average curves for each home for each season, the curves were put into groups via a *k-means* clustering analysis. *K-means* is an iterative process that attempts to partition each curve (the seasonal average curve – a 24-valued vector) into groups such that the sum of squares (Euclidean distance) from each group member to its group center is minimized [18]. After each step, the centers are recalculated and the curves are redistributed until the process converges. In this approach, as in most clustering analyses, the number of clusters is pre-specified, so prior knowledge of the number of groups is required. There are tests that attempt to determine the optimal number of clusters in the complete dataset, but different tests can give slightly different results. In this analysis, the *R* function *NbClust* was used to determine the optimal number of clusters [19]. The function *NbClust* performs 30 tests for determining the optimal number of clusters in the complete group of data. The function returns each of the tests optimal number of clusters, and the most-picked number of clusters was chosen as the number to use in the *k-means* clustering analysis. The test of the number of optimal clusters was performed for each season separately.

The temporal shape of the electricity use profile is important because even though residential customers typically pay a constant rate per kWh for electricity, the wholesale electric market price can fluctuate orders of magnitude in less than an hour. These dynamics are particularly observed during the summer peak in the Electric Reliability Council of Texas (ERCOT) grid – the electric grid that services most of Texas, including Austin. For a given season, the average value of the cluster’s electricity use can be defined by

Equation 3:

$$E_{value,z} = \frac{\sum^h (C_{z,h} \times SPP_h)}{kWh_z}, \quad (3)$$

where $E_{value,z}$ is the average value of cluster z 's electricity use (\$/kWh), $C_{z,h}$ is the average (non-normalized) electricity use for cluster z at hour h , SPP_h is the average price of wholesale market electricity at hour h , and kWh_z is the average total kWh of electricity used per day by cluster z . **Note that the homeowners are not subject to these prices – they all pay the same seasonal tiered rate structure** [20].

2.3. Binary regression analysis of clustering results

Once clusters of profiles were created for each season, survey data from the homes were used to see if there were any significant correlations between survey responses and cluster identification. A probit regression model (R , *probitmfx* [21]) was chosen to test the significance of any marginal effects of explanatory variables on cluster identification. A marginal effect of an explanatory variable, x , is the partial derivative, with respect to x , of the prediction function (in this case the normal cumulative probability density function) with all the other explanatory variables at their average value. Because this method is based on the partial derivative, the marginal effect can be greater than one even though the probability is between 0 and 1. The probit model is a binary classification model where the dependent variable (y_j) can only take on a value of 0 or 1. In this case, the model attempted to estimate the inclusion in a certain cluster (1 if included and 0 if not). The model is given by Equation 4:

$$Pr(y_j = 1) = \Phi(X_j\beta), \quad (4)$$

where Pr is the probability, y_j is the outcome (response) variable (0 or 1 for home j), Φ is the normal cumulative distribution function with mean 0 and standard deviation 1, X_j is the set of explanatory variables, and β is the set of fit coefficients (usually decided by maximum likelihood) [22].

The R function *probitmfx*, with robust standard errors, was chosen to build the model. This function is convenient because the coefficients (β) returned are the marginal effects of the explanatory variables (X_i) on the probability of y_i being equal to 1. Each β is interpreted as the effect of each variable at the average value of the others.

Explanatory variables in the model included: *workday* (whether or not someone worked from home for more than 20 hours per week during the day time, 59 total); *males* (the number of males living in the home); *females* (the number of females living in the home); *num_kids* (the number of children under 18 living in the home); *education* (the highest education level of the occupants); *income* (the income bracket of the home); *hours_tv* (the number of hours spent watching television per week); *num_computers* (the number of desktop computers, laptops, tablets, and other computing devices in the home); *vehicles* (the number of vehicles in the home); *quiz_total* (the number of questions correctly answered in the survey that referred to general energy and water knowledge); *p_v* (whether or not a home had a solar PV system, 63 total); *bayes* (how the homeowner answered a question on happiness [15]); *ev* (whether or not the home has an electric vehicle (EV), 26 total); *greenbuilt* (whether the home was a new green-built home, 78 total); *retrofit* (whether or

not the home has had any major retrofits, 8 total); *smart_therm* (whether or not the home has a smart thermostat, 26 total); *pv:greenbuilt* (the interaction term of being a new green-built home and having a solar PV system – most homes with solar PV systems were also green-built); and *ev:in_greenbuilt* (the interaction term of being a new green-built home and having an EV – most homes with EVs were also green-built). Figure 1 shows the statistical breakdown of the continuous variables. Figures 3 – 4 show histograms of the categorical variables, education and income, used in this analysis.

Table 1: Statistical breakdown of the continuous variables used in this analysis.

Variable	Mean	Median	Standard Deviation
males	1.2	1	0.75
females	1.3	1	0.78
num_kids	0.7	0	0.92
hours_tv	3.6	3	2.51
computers	3.1	3	1.60
vehicles	1.9	2	0.65
quiz_total	5.0	5	2.00

The model was validated by the Wald χ^2 statistic [23] and by checking the percentage of times that it could correctly predict the inclusion in a given cluster. The Wald χ^2 statistic checks if at least one of the explanatory variables (β) in the model is non-zero.

Figure 3: Histogram of the levels of education in the home.

Figure 4: Histogram of the levels of income in the home.

3. Results and Discussions

3.1. Creating average seasonal curves

Seasonal curves were created for each home for each of the four seasons. Figures 5 – 8 show box plots of the pre-normalized curves for all the homes with a distribution for each hour of the day. The graphs provide distributions of electric power demand for every hour of the day in each season for all the homes in this analysis. The seasons have varying levels of magnitude with the difference likely driven by residential air-conditioning use [24, 25], which is a function of climate, homeowner preference, home construction, and air-conditioner efficiency.

In all four figures, the solid line in the middle of the box plot refers to the median value for that hour for all the curves. The length from the median to the bottom of the box is the second quartile (25% of values) and the distance from the median to the top of the box is the third quartile. The length from the bottom of the box to the bottom whisker is the shorter of 1.5 times the inter-quartile range (top of the box to the bottom) or the distance from the bottom of the box to the lowest data point for that hour and the length to the top of the box to the top whisker is the shorter of 1.5 times the inter-quartile range or the distance from the top of the box to the highest data point for that hour. Any circles are data points that are beyond 1.5 times the inter-quartile range from the bottom or top of the box.

These results highlight the difference in residential inter-seasonal electricity use caused by weather-driven space conditioning loads. The seasons were then used to dissect the data further in search of inter-seasonal variations between homes' usage profiles.

Figure 5: Box-and-whisker plot of the raw (pre-normalization) curves for all homes for the summer (June 2013 – August 2013) season. The box-and-whisker shows the 4 quartiles of the data and the circles are data points above or below 1.5 times the interquartile range from either the top or bottom of the box.

Figure 6: Box-and-whisker plot of the raw (pre-normalization) curves for all homes for the fall (September 2013 – October 2013 + November 2012) season. The box-and-whisker shows the 4 quartiles of the data and the circles are data points above or below 1.5 times the interquartile range from either the top or bottom of the box.

Figure 7: Box-and-whisker plot of the raw (pre-normalization) curves for all homes for the winter (December 2012 – February 2013) season. The box-and-whisker shows the 4 quartiles of the data and the circles are data points above or below 1.5 times the interquartile range from either the top or bottom of the box.

Figure 8: Box-and-whisker plot of the raw (pre-normalization) curves for all homes for the spring (March 2013 – May 2013) season. The box-and-whisker shows the 4 quartiles of the data and the circles are data points above or below 1.5 times the interquartile range from either the top or bottom of the box.

3.2. Determining optimal distinctive clusters within seasons.

It was determined that each season has two clusters that represent seasonal household electricity demand. That is, every home, in each season falls in one of two groups. Figures 9 – Figure 12 show the normalized profiles clustered into groups as well as the cluster mean.

Figure 9: Plot showing (left and center) the breakout of summer clusters including the cluster averages in red and (on the right) both cluster averages in the same plot with the normalized ERCOT SPP price for the same time period.

Summer cluster 1 (Figure 9 left) included homes that (on average) tended to have a wide range of early morning usage that reduces until about 07:00 and then increases to an afternoon peak around 19:00, and then reduces. Cluster 2 (Figure 9 center) included homes that (on average) start with

high usage which decreases until about 10:00 and then increases until about 19:00 at which the homes maintained an average level of power draw for the rest of the day. Summer cluster 1 is tighter than cluster 2 with an average within cluster sum of squares of 0.31 and 0.50, respectfully. Figure 9 (right) also includes the average summer cluster profiles on the same plot as the normalized average ERCOT SPP price. The normalized summer average ERCOT SPP price curve peaked at 17:00. The ERCOT SPP price peaks at relatively high prices in the afternoon summer hours with 16 of 24 hours at a price less than 25% of the peak price. Even though individual homes could be peaking coincident with the ERCOT peak, when the summer ERCOT SPP price peaks, the cluster 1 average is at 91% of its maximum and the cluster 2 average is at 68% of its maximum. Based on ERCOT's SPP, the average summer value (Equation 3) of cluster 1's electricity use is about 10% higher than from cluster 2 (0.0528 verses 0.0480 \$/kWh).

The fall (Figure 10) ERCOT SPP price peaks are less severe than the summer time peak probably due to less residential air-conditioning demand and the greater availability of low-cost wind generation on the grid [26]. Similar to the the summer clusters, when the fall ERCOT SPP price peaks, the cluster 1 average is at 88% of its maximum and the cluster 2 average is at 55% of its maximum. Based on ERCOT's SPP, the average fall value (Equation 3) of cluster 1's electricity use is about 5% higher than from cluster 2 (0.0333 verses 0.0317 \$/kWh).

The winter (Figure 11) ERCOT SPP price peaks is again less severe than the summer time peak and the early morning peak is probably explained by electric resistance heating or electric heat pumps in many homes in ER-

Figure 10: Plot showing (left and center) the breakout of fall clusters including the cluster averages in red and (on the right) both cluster averages in the same plot with the normalized ERCOT SPP price for the same time period.

Figure 11: Plot showing (left and center) the breakout of winter clusters including the cluster averages in red and (on the right) both cluster averages in the same plot with the normalized ERCOT SPP price for the same time period.

COT [27], and the fact that the relative difference between the morning and evening usage is much less in the winter. Because the mild differences in the winter ERCOT SPP price, the average value of cluster 1 and cluster 2's electricity use was essentially the same.

Figure 12: Plot showing (left and center) the breakout of spring clusters including the cluster averages in red and (on the right) both cluster averages in the same plot with the normalized ERCOT SPP price for the same time period.

The spring peaks are probably a result of a combination of heating and cooling putting stress on the grid as the season transitions from winter to summer. Based on ERCOT's SPP, the average summer value (Equation 3) of cluster 1's electricity use is about 4% higher than from cluster 2 (0.0296 verses 0.0286 \$/kWh). The results from the clustering analysis for all seasons

are summarized in Table 2.

Table 2: Clustering analysis results showing tightness of clusters and the average seasonal value of electricity used by each cluster.

Season	Cluster	Tightness	Cost(\$/kWh)	Hour of Peak
Summer	1	0.31	0.0528	19:00
	2	0.5	0.0480	19:00
Fall	1	0.35	0.0333	18:00
	2	0.46	0.0317	21:00
Winter	1	0.49	0.0274	20:00
	2	0.81	0.0275	21:00
Spring	1	0.52	0.0296	19:00
	2	0.74	0.0286	23:00

3.3. Binary regression analysis of clustering results

Binomial probit regression was used to determine if there were any significant correlations affecting the probability of a home being in a certain cluster. The correlation matrix of the explanatory variables themselves did not find any overwhelming correlation between any two variables – this correlation matrix can be seen in Table A.7 of Appendix A. The output from the analysis of the effect of survey variables on the prediction of a home being in cluster 1 during the summer season is given in Table 3. Because there are only two outcomes (cluster 1 or cluster 2) the regression output is the same for the effect of variables prediction of a home being in cluster 2 except the

signs on the coefficient estimates are reversed and interpreted as having the opposite effect.

The variables *workday*, *hours_tv*, *computers*, *retrofit*, and the interaction term *pv:greenbuilt* are all significant using a confidence interval of 95% and the variables *females*, *education*, and *income* are significant using a confidence interval of 90%. The coefficient estimates are calculated as the direct marginal effect (dF/dx) of the explanatory variable on the probability of a home being in cluster 1 with all other explanatory variables at their average value. Thus, if a home contains a person that regularly works from home (*workday*), that home is 55% more likely to be in cluster 1. Each additional hour of television watched per week (*hours_tv*) corresponds to a 10% greater chance of being in cluster 1 (this variable has also been shown to be significant in macro-level studies of electricity use [5]). Each additional computing device (*computers*) in the home leads to a 13% increase in the probability of being in cluster 1. Also, if the home has had recent energy retrofits, it is 42% more likely to be in cluster 1. Lastly, if the home has solar PV *and* is green-built it is 64% more likely to be in cluster 1. The variables *females*, *education*, and *income* were negatively correlated with being in cluster 1, although with less certainty. The χ^2 value of 30.3 indicates that the model is significant compared to the null model (all $\beta = 0$). Because cluster 1 consumes relatively more electricity when prices are higher (see Figure 9), the regression results indicate that people who work from home are more likely to be negatively impacted by time-of-use or real-time pricing schemes (all else equal), which are often more prevalent in the summer season.

The model was able to predict the correct summer cluster about 77% of

Table 3: Probit regression results for summer profiles showing marginal effects (dF/dx) of each explanatory variable (β) on the probability of being in cluster 1, at the average of all the others. Summer cluster 1 is the left box of Figure 9.

Explanatory Variable	Coefficient (dF/dx)	Robust Std. Error	z-value	Two-tailed P-test	95% Confidence Interval	
workday	0.55	0.11	5.26	0.00	(0.34 , 0.76)	***
males	-0.24	0.19	-1.26	0.21	(-0.63 , 0.14)	
females	-0.37	0.21	-1.77	0.08	(-0.78 , 0.05)	o
num_kids	0.19	0.20	0.91	0.36	(-0.22 , 0.60)	
education	-0.23	0.12	-1.91	0.06	(-0.47 , 0.01)	o
income	-0.11	0.06	-1.90	0.06	(-0.22 , 0.01)	o
hours_tv	0.10	0.03	3.41	0.00	(0.04 , 0.16)	**
num_computers	0.13	0.05	2.81	0.01	(0.04 , 0.23)	**
vehicles	0.13	0.12	1.09	0.28	(-0.11 , 0.37)	
quiz_total	-0.02	0.03	-0.49	0.62	(-0.08 , 0.05)	
pv	-0.43	0.40	-1.08	0.28	(-1.23 , 0.37)	
bayes	0.25	0.19	1.29	0.20	(-0.14 , 0.63)	
ev	-0.50	0.37	-1.36	0.18	(-1.24 , 0.23)	
greenbuilt	0.27	0.23	1.17	0.24	(-0.19 , 0.72)	
retrofit	0.43	0.11	3.76	0.00	(0.20 , 0.65)	***
smart_therm	0.16	0.19	0.83	0.41	(-0.22 , 0.53)	
pv:greenbuilt	0.65	0.33	1.99	0.05	(0.00 , 1.30)	*
ev:greenbuilt	0.02	0.55	0.03	0.98	(-1.08 , 1.11)	

$\chi^2 = 30.3$, $n = 103$, Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 'o' 0.1 ' ' 1

the time. Regression texts typically suggest methods for removing explanatory variables based on the expected sign of the coefficient, significance (z or t -scores), and knowledge of the underlying data and population [22]. Little was known about how certain variables would intuitively affect which cluster a home was likely to be in with the exception of the variable *workday*. Thus, all variables were left in the model. Note that if a home had PV ($n = 63$), the direct impact would not show up in the curve, i.e. the clustered curves were the temporal power demand consumed within the home and any exterior appliances such as lighting and the air-conditioner compressor. The charging of EVs ($n = 26$) would show up in the homes' temporal power demand, however if homeowners plugged in their EVs when they arrived home in the afternoon when the air-conditioner is typically on (in most of the seasons in Austin), it would show up as an increased normalized deference between the afternoon and midday usage. Another reason that EVs might not show up in the regression analysis is that people might not always be consistent with their starting charge times or duration [28–30].

The results for the fall, winter, and spring probit regression models are given in Tables 4 – 6.

All of the models found *workday* to be significant, at least to the 90% confidence interval and associated with cluster 1 of each season. Cluster 1 for the summer, fall, and spring seasons is marked with an earlier daily rise in energy use than cluster 2. The variable *hours_tv* was positively correlated with the summer, fall, and spring season's cluster 1 and *education* was negatively correlated with the same three seasons' cluster 1. The variables *retrofit* and *computers* were positively associated with summer and spring's cluster 1

Table 4: Probit regression results for fall profiles showing marginal effects (dF/dx) of each explanatory variable (β) on the probability of being in cluster 1, at the average of all the others. The model was able to predict the correct summer cluster only about 65% of the time. Fall cluster 1 is the left box of Figure 10.

Explanatory Variable	Coefficient (dF/dx)	Robust Std. Error	z-value	Two-tailed P-test	95% Confidence Interval	
workday	0.35	0.12	2.89	0.00	(0.11 , 0.59)	**
males	-0.14	0.16	-0.91	0.37	(-0.45 , 0.17)	
females	-0.17	0.16	-1.05	0.30	(-0.49 , 0.15)	
num_kids	0.05	0.16	0.33	0.74	(-0.27 , 0.38)	
education	-0.07	0.10	-0.65	0.52	(-0.28 , 0.14)	
income	-0.06	0.05	-1.36	0.18	(-0.16 , 0.03)	
hours_tv	0.05	0.02	1.95	0.05	(0.00 , 0.10)	o
computers	0.07	0.04	1.73	0.09	(-0.01 , 0.15)	o
vehicles	0.08	0.12	0.67	0.51	(-0.16 , 0.32)	
quiz_total	-0.02	0.03	-0.62	0.53	(-0.08 , 0.04)	
pv	0.09	0.40	0.22	0.83	(-0.71 , 0.89)	
bayes	0.10	0.16	0.61	0.55	(-0.23 , 0.43)	
ev	-0.37	0.41	-0.91	0.37	(-1.18 , 0.44)	
greenbuilt	0.19	0.21	0.89	0.37	(-0.23 , 0.60)	
retrofit	0.22	0.15	1.43	0.16	(-0.09 , 0.52)	
smart_therm	0.14	0.17	0.86	0.39	(-0.19 , 0.48)	
pv:greenbuilt	0.25	0.42	0.60	0.55	(-0.59 , 1.10)	
ev:greenbuilt	-0.14	0.47	-0.31	0.76	(-1.08 , 0.80)	

$\chi^2 = 29.1$, $n = 103$, Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 'o' 0.1 ' ' 1

Table 5: Probit regression results for winter profiles showing marginal effects (dF/dx) of each explanatory variable (β) on the probability of being in cluster 1, at the average of all the others. The model was able to correctly predict cluster identification about 82% of the time. Winter cluster 1 is the left box of Figure 11.

Explanatory Variable	Coefficient (dF/dx)	Robust Std. Error	z-value	Two-tailed P-test	95% Confidence Interval	
workday	0.14	0.08	1.67	0.10	(-0.03 , 0.30)	o
males	0.11	0.11	0.97	0.33	(-0.12 , 0.34)	
females	0.12	0.09	1.31	0.19	(-0.06 , 0.29)	
num_kids	-0.09	0.13	-0.69	0.49	(-0.34 , 0.17)	
education	-0.15	0.08	-1.80	0.08	(-0.32 , 0.02)	o
income	-0.04	0.04	-1.02	0.31	(-0.11 , 0.04)	
hours_tv	0.00	0.01	-0.20	0.84	(-0.03 , 0.03)	
computers	0.01	0.03	0.41	0.69	(-0.05 , 0.08)	
vehicles	-0.08	0.08	-1.09	0.28	(-0.24 , 0.07)	
quiz_total	0.03	0.03	1.23	0.22	(-0.02 , 0.09)	
pv	0.94	0.06	16.33	< 2.20E-16	(0.83 , 1.06)	***
bayes	-0.10	0.08	-1.22	0.22	(-0.27 , 0.06)	
ev	-1.00	0.00	-574.48	< 2.20E-16	(-1.00 , -0.99)	***
greenbuilt	0.00	0.09	-0.01	0.99	(-0.18 , 0.18)	
retrofit	-0.01	0.12	-0.12	0.91	(-0.25 , 0.22)	
smart_therm	-0.38	0.18	-2.10	0.04	(-0.75 , -0.02)	*
pv:greenbuilt	-0.84	0.11	-7.93	0.00	(-1.06 , -0.63)	***
ev:greenbuilt	0.51	0.15	3.50	0.00	(0.22 , 0.80)	***

$\chi^2 = 32.98$, n = 103, Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6: Probit regression results for spring profiles showing marginal effects (dF/dx) of each explanatory variable (β) on the probability of being in cluster 1, at the average of all the others. The model was able to correctly predict cluster identification about 79% of the time. Spring cluster 1 is the left box of Figure 12.

Explanatory Variable	Coefficient (dF/dx)	Robust Std. Error	z-value	Two-tailed P-test	95% Confidence Interval	
workday	0.28	0.12	2.40	0.02	(0.05 , 0.51)	*
males	0.14	0.16	0.90	0.37	(-0.18 , 0.47)	
females	0.21	0.17	1.26	0.21	(-0.12 , 0.55)	
num_kids	-0.15	0.18	-0.82	0.41	(-0.51 , 0.21)	
education	-0.23	0.11	-2.10	0.04	(-0.45 , -0.01)	*
income	-0.11	0.05	-2.26	0.03	(-0.21 , -0.01)	*
hours_tv	0.05	0.03	1.99	0.05	(0.00 , 0.10)	o
computers	0.04	0.05	0.76	0.45	(-0.07 , 0.14)	
vehicles	0.00	0.11	-0.03	0.98	(-0.23 , 0.22)	
quiz_total	-0.05	0.04	-1.28	0.20	(-0.12 , 0.03)	
pv	-0.40	0.27	-1.47	0.15	(-0.94 , 0.14)	
bayes	0.28	0.16	1.70	0.09	(-0.05 , 0.60)	o
ev	-0.40	0.37	-1.09	0.28	(-1.14 , 0.34)	
greenbuilt	0.28	0.22	1.25	0.21	(-0.17 , 0.73)	
retrofit	0.24	0.13	1.84	0.07	(-0.02 , 0.49)	o
smart_therm	-0.01	0.25	-0.02	0.98	(-0.50 , 0.49)	
pv:greenbuilt	0.51	0.35	1.46	0.15	(-0.19 , 1.21)	
ev:greenbuilt	-0.22	0.44	-0.49	0.62	(-1.10 , 0.66)	

$\chi^2 = 31.9$, $n = 103$, Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 'o' 0.1 ' ' 1

while *income* was negatively correlated with the same season’s cluster 1.

Variables such as *workday*, *hours_tv*, and *computers* are likely associated with higher relative electricity use during the day based on the behavioral impacts of being at home and the use of multiple pieces of technology. Plug loads are increasing as a share of the total energy use in the home [31], and these results indicate that this extra electricity draw might be during times of higher wholesale electricity prices. As expected, the variable *workday* was strongest during the summer as space cooling during the day would be desired by someone at home during the hot summers.

Surprisingly, the variable *retrofit* was positively associated with cluster 1 in both the summer and spring seasons, this might indicate either some sort of rebound effect in not only magnitude of energy use, but in the temporal aspect of that energy use as well. More study is needed in both the classical rebound effect and the temporal aspect that this analysis might have observed.

Residential customers typically pay a flat fee, on a \$/kWh basis, for electricity. While some pricing structures are tiered based on total consumption and some utilities offer time-of-use rates, currently most rates in Texas are not dynamically tied to market conditions. Electric utilities act as a buffer for the market and attempt to charge a price that, over a fiscal year, will account for market fluctuations. There seems to be a distinct group of customers that typically use electricity when the price is higher, possibly narrowing the utilities’ profit margins. However, homeowners do not typically have a financial reason to care about this situation as they are not subject to the temporal fluctuations in price, but might indirectly be so when utilities consider rates

cases (changing electric pricing structures). This analysis shows that there are markers of homes that typically use higher priced electricity and thus utilities might be able to target such customers with incentives that would reduce or shift electricity usage to less expensive times, possibly increasing profit margins and keeping rates lower overall.

4. Conclusions

This analysis employed a data-driven approach to 1) determine the shape of seasonally-resolved residential demand profiles, 2) determine the optimal number of normalized representative residential electricity use profiles within each season, and 3) draw correlations to the different profiles based on survey data for over 100 homes in Austin, TX. This analysis determined that, for homes in a southern U.S. location, there are two main groups of temporal profiles, representing residential electricity demand in each season. Temporal and magnitude differences in the summer profiles are significant with one group typically demanding more power during higher wholesale electricity price times than the other. Finally, probit regression analysis determined that explanatory variables such as whether someone worked at home, the number of hours of television watched per week, and education were significant determiners of inclusion in a given cluster.

This analysis could also serve as a starting point for utilities looking to reduce electricity use during peak times. The drivers of the temporal shape of residential use could allow for policies that target load shifting and efficiency upgrade rebates to homes that would benefit the utility the most, such as homes with persons who work from home. This work and the methods

herein could be used to predict how changing demographics of neighborhoods could influence local distribution grid conditions. Also, further work is needed to assess the possible temporal rebound effect of residential energy retrofits. Policy interventions (or exceptions) for lower income households in areas that switch to time-of-use or real-time electricity price structures might be necessary to ensure that lower income household’s electricity bills are not disproportionately increased. These results might also inform local or regional planning organization’s policies on incentives for likely household demographic targets or changes with a focus towards reducing volatility in an aggregate electric demand curve, which have implications for distribution and transmission line planning.

5. Acknowledgments

The authors would like to thank the US Department of Energy, the Doris Duke Charitable Foundation, Austin Energy, the Texas Advanced Computing Center, and Pecan Street Inc. for their direct and in-kind support. The authors would also like to thank the study participants and the Mueller neighborhood groups for their support. This material is partly based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1110007. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

6. Disclosure of Affiliation

In addition to research work on topics generally related to energy systems at the University of Texas at Austin, some of the authors are equity partners in IdeaSmiths LLC, which consults on topics in the same areas of interests. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

Appendix A. Supplementary tables

Table A.7: Correlation matrix of the explanatory variables used in the probit regression analysis of the clusters.

	greenbuilt	workday	males	females	num_kids	education	income	hours.tv	computers	vehicles	quiz_total	pv	bayes	ev	retrofit	smart_therm
greenbuilt	1.00	0.01	-0.03	0.11	0.12	0.06	0.13	0.09	0.17	-0.12	-0.06	0.43	0.00	0.07	0.00	-0.56
workday	0.01	1.00	0.10	0.04	0.01	0.04	-0.04	0.26	0.08	0.01	0.05	0.04	0.06	0.00	0.03	-0.18
males	-0.03	0.10	1.00	-0.05	0.60	0.20	0.31	-0.15	0.24	0.17	0.18	-0.03	0.05	-0.13	0.00	0.05
females	0.11	0.04	-0.05	1.00	0.65	0.07	-0.03	-0.07	0.28	0.29	-0.13	0.04	-0.06	-0.12	0.07	-0.12
num_kids	0.12	0.01	0.60	0.65	1.00	0.12	0.12	-0.26	0.28	0.10	0.04	0.04	0.03	-0.25	-0.01	-0.01
education	0.06	0.04	0.20	0.07	0.12	1.00	0.23	-0.18	0.24	0.03	0.14	0.05	-0.03	0.03	0.11	-0.05
income	0.13	-0.04	0.31	-0.03	0.12	0.23	1.00	-0.02	0.23	0.23	0.00	-0.06	-0.19	0.02	0.05	-0.18
hours.tv	0.09	0.26	-0.15	-0.07	-0.26	-0.18	-0.02	1.00	-0.10	0.04	-0.09	-0.08	-0.08	0.00	-0.04	-0.19
computers	0.17	0.08	0.24	0.28	0.28	0.24	0.23	-0.10	1.00	0.25	-0.06	0.06	0.04	0.10	0.05	-0.17
vehicles	-0.12	0.01	0.17	0.29	0.10	0.03	0.23	0.04	0.25	1.00	-0.01	-0.10	-0.10	0.09	0.14	-0.05
quiz_total	-0.06	0.05	0.18	-0.13	0.04	0.14	0.00	-0.09	-0.06	-0.01	1.00	0.32	0.35	-0.01	0.06	-0.01
pv	0.43	0.04	-0.03	0.04	0.04	0.05	-0.06	-0.08	0.06	-0.10	0.32	1.00	0.61	0.33	0.01	-0.36
bayes	0.00	0.06	0.05	-0.06	0.03	-0.03	-0.19	-0.08	0.04	-0.10	0.35	0.61	1.00	0.24	0.09	0.02
ev	0.07	0.00	-0.13	-0.12	-0.25	0.03	0.02	0.00	0.10	0.09	-0.01	0.33	0.24	1.00	0.25	-0.13
retrofit	0.00	0.03	0.00	0.07	-0.01	0.11	0.05	-0.04	0.05	0.14	0.06	0.01	0.09	0.25	1.00	0.00
smart_therm	-0.56	-0.18	0.05	-0.12	-0.01	-0.05	-0.18	-0.19	-0.17	-0.05	-0.01	-0.36	0.02	-0.13	0.00	1.00

References

- [1] EIA, . Annual Energy Review 2010. Tech. Rep.; Energy Information Administration, U.S. Department of Energy; 2011. URL: <http://www.eia.gov/totalenergy/>.
- [2] Doggett, T.. Overcoming Barriers to Smart Grids & New Energy Services Electric Reliability Council of Texas. In: UT Smart Grid Conference. Austin; 2011,.
- [3] Katzenstein, W., Apt, J.. Air Emissions Due To Wind And Solar Power. Environmental Science & Technology 2009;43(2):253–258. URL: <http://www.ncbi.nlm.nih.gov/pubmed/22877159><http://www.ncbi.nlm.nih.gov/pubmed/19238948><http://pubs.acs.org/doi/abs/10.1021/es801437t>. doi:doi:10.1021/es801437t.
- [4] Meehan, C.M., King, C.W., Garrison, J.B., Webber, M.E.. The Total Impact of Wind Energy Variability on Fossil Fuel Emission Rates in Texas. In: 31st USAEE/IAEE North American Conference. USAEE; 2012, p. 1–17. URL: <http://www.usaee.org/usaee2012/submissions/ExtendedAbs/ExtendedAbstractMeehan.doc>.
- [5] Sanquist, T.F., Orr, H., Shui, B., Bittner, A.C.. Lifestyle factors in U.S. residential electricity consumption. Energy Policy 2012;42(0):354–364. URL: <http://www.sciencedirect.com/science/article/pii/S0301421511009906><http://linkinghub.elsevier.com/retrieve/pii/S0301421511009906>. doi:doi:10.1016/j.enpol.2011.11.092.

- [6] Mokhtarian, P.L., Varma, K.V.. The trade-off between trips and distance traveled in analyzing the emissions impacts of center-based telecommuting. *Transportation Research Part D: Transport and Environment* 1998;3(6):419–428. URL: <http://linkinghub.elsevier.com/retrieve/pii/S1361920998000182>. doi:doi:10.1016/S1361-9209(98)00018-2.
- [7] Warren Liao, T., Liao, T.W.. Clustering of time series data – a survey. *Pattern Recognition* 2005;38(11):1857–1874. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0031320305001305><http://www.sciencedirect.com/science/article/pii/S0031320305001305>. doi:doi:10.1016/j.patcog.2005.01.025.
- [8] Reddy, T.A.. Chapter 8: Classification and Clustering Methods. In: *Applied Data Analysis and Modeling for Energy Engineers and Scientists*. 2011, p. 231–251. URL: <http://www.springer.com/energy/energy+efficiency/book/978-1-4419-9612-1>.
- [9] Chicco, G.. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy* 2012;42(1):68–80. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0360544211008565><http://www.sciencedirect.com/science/article/pii/S0360544211008565>. doi:doi:10.1016/j.energy.2011.12.031.
- [10] Kim, Y.I., Ko, J.M., Choi, S.H.. Methods for generating TLPs (typical load profiles) for smart grid-based energy programs. In:

- 2011 IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG). IEEE. ISBN 978-1-4244-9893-2; 2011, p. 1–6. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5953331>. doi:doi:10.1109/CIASG.2011.5953331.
- [11] Kwac, J., Flora, J., Rajagopal, R.. Household Energy Consumption Segmentation Using Hourly Data. IEEE Transactions on Smart Grid 2014;5(1):420–430. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6693793>. doi:doi:10.1109/TSG.2013.2278477.
- [12] Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., Kolehmainen, M.. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Applied Energy 2010;87(11):3538–3545. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0306261910001832><http://www.sciencedirect.com/science/article/pii/S0306261910001832>. doi:doi:10.1016/j.apenergy.2010.05.015.
- [13] Zhou, K.l., Yang, S.l., Shen, C.. A review of electric load classification in smart grid environment. Renewable and Sustainable Energy Reviews 2013;24:103–110. URL: <http://www.sciencedirect.com/science/article/pii/S1364032113001822>. doi:doi:10.1016/j.rser.2013.03.023.
- [14] Panapakidis, I.P., Alexiadis, M.C., Papagiannis, G.K.. Electricity customer characterization based on different representative load curves. In: 2012 9th International Conference on the European Energy Market. IEEE. ISBN 978-1-4673-0833-5; 2012, p.

- 1–8. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6254761>. doi:doi:10.1109/EEM.2012.6254761.
- [15] Rhodes, J.D., Upshaw, C.R., Harris, C.B., Meehan, C.M., Walling, D.A., Navrátil, P.A., et al. Experimental and data collection methods for a large-scale smart grid deployment: Methods and first results. *Energy* 2014;65:462–471. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0360544213009663>. doi:doi:10.1016/j.energy.2013.11.004.
- [16] The Texas Advanced Computing Center. 2013. URL: <http://www.tacc.utexas.edu/home>.
- [17] What is Pecan Street. 2014. URL: <http://www.pecanstreet.org/about/what-is-pecan-street-inc/>.
- [18] Hothorn, T., Everitt, B.S.. *A Handbook of Statistical Analyses Using R*, Second Edition. Chapman and Hall/CRC; 2009. ISBN 1420079336. URL: <http://www.amazon.com/Handbook-Statistical-Analyses-Second-Edition/dp/1420079336>.
- [19] Package NBClust. 2014. URL: <http://cran.r-project.org/web/packages/NbClust/NbClust.pdf>.
- [20] Austin Energy, . Data:F85ed7ef-e7bf-4c6c-af37-0a082120125f — Open Energy Information. 2014. URL: http://en.openei.org/w/index.php?title=Data:F85ed7ef-e7bf-4c6c-af37-0a082120125f&urdbView=true#tab=3__Energy.

- [21] Fernihough, A.. CRAN - Package mfx. 2013. URL: <http://cran.r-project.org/web/packages/mfx/index.html>.
- [22] Gelman, A., Hill, J.. Data Analysis Using Regression and Multilevel/Hierarchical Models; vol. 625 of *Analytical methods for social research*. 11 ed.; Cambridge: Cambridge University Press; 2006. ISBN 9780511790942. URL: <http://books.google.com/books?id=c9xLKzZWoz4C>. doi:doi:10.1017/CBO9780511790942.
- [23] Engle, R.F.. Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of Econometrics* 1984;2:775–826. URL: <http://www.sciencedirect.com/science/article/pii/S1573441284020055>. doi:doi:http://dx.doi.org/10.1016/S1573-4412(84)02005-5.
- [24] Rhodes, J.D., Stephens, B., Webber, M.E.. Using energy audits to investigate the impacts of common air-conditioning design and installation issues on peak power demand and energy consumption in Austin, Texas. *Energy and Buildings* 2011;43(11):3271–3278. URL: <http://www.sciencedirect.com/science/article/pii/S0378778811003823>. doi:doi:10.1016/j.enbuild.2011.08.032.
- [25] Rhodes, J.D., Stephens, B., Webber, M.E.. Energy Audit Analysis of Residential Air-Conditioning Systems in Austin, Texas. *ASHRAE Transactions* 2012;118(1):143–150.
- [26] Srivastava, A.K., Pineda, R.L., Mandal, P.. Wind power generation impact on electricity price in ERCOT. In: 2012 IEEE Power and En-

- ergy Society General Meeting. IEEE. ISBN 978-1-4673-2729-9; 2012, p. 1–7. URL: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6345433>. doi:doi:10.1109/PESGM.2012.6345433.
- [27] Residential Energy Consumption Survey (RECS) - Energy Information Administration. 2013. URL: <http://www.eia.gov/consumption/residential/>.
- [28] Harris, C.B., Webber, M.E.. An empirically-validated methodology to simulate electricity demand for electric vehicle charging. *Applied Energy* 2014;126:172–181. URL: <http://linkinghub.elsevier.com/retrieve/pii/S0306261914003183>. doi:doi:10.1016/j.apenergy.2014.03.078.
- [29] Harris, C.B.. Personal Correspondence with Chioke B. Harris, PhD; 2014.
- [30] McCracken, B., Fisher, G., Hersey, A., Shead, J., Sokolic, K.. Residential EV Charging Diverging Significantly from Models; 2013.
- [31] Heating and cooling no longer majority of U.S. home energy use – U.S. Energy Information Administration (EIA). 2013. URL: <http://www.eia.gov/todayinenergy/detail.cfm?id=10271>.