Load Shape Clustering Using Residential Smart Meter Data: a Technical Memorandum

Authors:

Ling Jin[1], Anna Spurlock[1], Sam Borgeson[2], Daniel Fredman[3], Liesel Hans[4], Siddarth Patel[5], Annika Todd[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA
[2]Convergence Data Analytics LLC, Oakland, CA
[3]University of Vermont, Burlington, VT
[4]City of Fort Collins, CO
[5]Stanford University, CA

Energy Analysis and Environmental Impacts Division
Lawrence Berkeley National Laboratory

Sustainable Energy Systems
Electricity Markets and Policy

September 2016

This work was supported by the U.S Department of Energy under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

COPYRIGHT NOTICE

Anna Spurlock
Lawrence Berkeley
National Lab
CASpurlock@lbl.gov

Annika Todd
Lawrence Berkeley
National Lab
atodd@lbl.gov

Ling Jin
Lawrence Berkeley
National Lab
ljin@lbl.gov

This document was final as of September 2016. If referenced, it should be cited as:
Jin, L., C. A. Spurlock, S. Borgeson, D. Fredman, L. Hans, S. Patel, and A. Todd. 2016. *Load Shape Clustering Using Residential Smart Meter Data : a Technical Memorandum*. Lawrence Berkeley National Laboratory.

Table of Contents

List of Figures

Executive Summary

The rise of Advanced Metering Infrastructure has enabled large volumes of electricity consumption data to be captured at an hourly frequency or even higher. A thread of research has demonstrated methods for coupling this fast growing data stream with data mining techniques such as cluster analysis for categorization of electricity load patterns. In past research on residential customers, such categorization has usually been conducted on aggregated load data, partly due to large variability exhibited within and across customers. However, in the context of demand response and efficiency programs, load patterns of individual customers and their daily and inter-daily variability directly relate to each customer's ability to respond to program incentives.

This document is a technical memorandum of application of an innovative clustering technique to individual customers' daily load data resolved at the hourly level across a large sample of residential customers over a full year period. An additional innovation of our work is that we focus our analysis on the timing of discretionary[1] electricity usage in particular, as opposed to total electricity use. We document the innovations and hyperparameter selection in the clustering process specific to our residential smart meter dataset and derive a diverse set of archetypical discretionary loadshapes.

While typically utilities and system operators focus on the aggregate residential load shape, application of this improved clustering method will shed light on the considerable heterogeneity and variability across days and customers. In the future, more behavioral features associated with household consumption schedules and variability can be extracted based on our results and can be used in future studies to segment customers for better program targeting and designing tailored recruitment strategies.

[1] We further define what we mean by "discretionary" later in the paper. In essence it is the usage above the daily minimum hourly usage, which is used as a proxy for the house's base load energy consumption.

1 Introduction

With the rise of Advanced Metering Infrastructure (AMI) in the past decade, large volumes of electricity consumption data can now be captured, stored and reported at 5 to 60-minute intervals. Such high-resolution consumption data can provide both utilities and customers with new insights on end-use and behavioral patterns, facilitating load planning and forecasting, demand response management, time-of-use tariff design, and electricity settlement (Moslehi and Kumar 2010; Farhangi 2010; Hong 2011; Zhou et al. 2013). This fast growing stream of data, coupled with innovative data analytics, holds the potential to change the landscape of the traditional electric utility industry by supporting intelligent power grids and smart energy management (Zhou et al. 2016).

Research using data mining techniques has started to emerge in order to segment customers based on daily patterns of hourly electricity consumption. This is referred to as "load profiling" or "load profile classes" (e.g. Wang 2015; McLoughlin 2015). As reviewed in Chicco (2012), a number of clustering techniques, such as k-means, follow the leader, self-organizing maps, etc., have been applied to whole-building load data to construct load profiles for non-residential (industrial and commercial) customers.

For residential customers, load profile classes are usually constructed based on aggregating the consumption patterns (e.g. using seasonal averages in Rhodes et al. 2014). As Chicco (2012) points out, residential customers are generally not treated as individual entities when conducting load pattern categorization because: (1) individual consumption patterns vary widely and are unpredictable, and (2) system feeders, the focus of demand characterization and prediction for utilities, supply aggregated loads. However, in the context of demand response (DR) and efficiency programs, individual load patterns directly relate to each customer's ability to respond to program incentives. For example, household activity levels during the system peak are relevant for their demand response potential, and the variability of daily load patterns over time may reflect the flexibility in household consumption schedules. By clustering individual load data, Kwac et al. (2014) found that although two homes might have the same average profiles, the "information entropy," or diversity of the two homes' load profiles from one day to the next, could vary significantly (e.g., one home could vary energy use patterns day to day, while the other home may have a more set usage pattern). McLoughlin et al. (2015) differentiated customer profile classes by their day-to-day usage patterns. This type of understanding of load patterns and their determinants at the level of individual residential customers could potentially lead to more effective program targeting and engagement, more precise prediction of demand response potential, more realistic grid planning, and more robust energy modeling.

The remaining sections of the report provide a detailed description of the improvement and application of the adaptive kmeans method originally developed by Kwac et al. (2014), including data preprocessing, hyperparameter selection, post-clustering process, and results.

2 Clustering Method

We cluster daily usage patterns based on hourly consumption data collected from a summer peaking utility in California. The data consist of over 30 million daily load profiles from approximately 100,000 households, which were measured between June 1st 2011 to May 31st 2012. Representative load shapes (hereafter referred to as "dictionary load shapes") are identified as cluster centers using a subset of daily load profiles (100,000 as suggested by Kwac et al. 2014) with the following steps: load data preprocessing, load clustering, post-cluster processing. Finally, ~30 million daily load profiles from the whole dataset are each assigned to the closest dictionary shape. The clustering method employed here builds upon the method developed in Kwac et al. (2014) and is illustrated in Figure 1 and described in detail below.

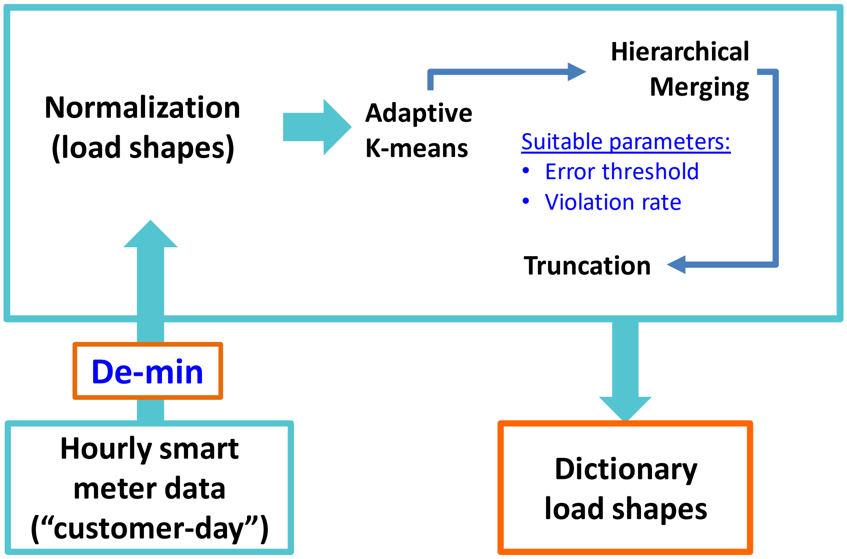

Figure 1 Flow chart illustrating derivation of representative load shapes.

2.1 Load data preprocessing

In Kwac et al., daily usage data with missing hours (0.3% of the data) or with very small power demand are ignored in populating their clustering. The cutoff of average power demand used to determine those that should be ignored is set to be 0.2kW, which corresponds to the 6% quantile. We apply the same preprocessing criteria, and in our case after data cleaning, 32,611,421 daily load shapes (94% of the raw data) are remaining. We then normalized each daily load shape observations by their respective daily total consumption so the area under each single customer 24-hour load shape is 1.

Our interests lie in characterizing discretionary consumption and in avoiding the flattening of normalized shapes with high baseload consumption. We therefore innovate beyond Kwac et al., and "de-min" load profiles prior to normalization and subsequent processing. Specifically, the daily minimum electricity usage is subtracted from each hour of each customer-day profile. This modified procedure has two advantages:

A household's "*discretionary*" usage captures the electricity consumption resulting from active residential behavior (e.g., lighting, air conditioning, computer equipment, entertainment, dishwashers, laundry equipment). We innovate beyond Kwac et al. (2014) to isolate only discretionary consumption by "de-minning" the load profiles prior to normalization. Specifically, the daily minimum electricity usage is subtracted from each hour of that day within each household-day profile. The object to be clustered is therefore defined to be this "de-minned" and normalized profile of discretionary daily usage. This procedure has two advantages: first, from a conceptual perspective daily minimum electricity usage serves as a proxy for "baseload" so this procedure allows us to isolate a household's variable or discretionary usage from their baseload. After normalization, a load shape essentially represents a sequence of each hour's proportional contribution to that day's total discretionary usage, and dictionary load shapes can be interpreted in terms of the overall patterns in timing of higher and lower discretionary use.

The second advantage to this "de-minning" process is that it alleviates the distortion of consumption profiles that occurs during normalization when using the total daily electricity load for each household. In particular, a load profile with high baseload tends to be flattened when total hourly usage is divided by daily total consumption in the normalization step. To demonstrate this, the top row of Figure 2 illustrates two load shapes with the same discretionary consumption schedules but different baseloads, and the bottom row shows those same shapes after normalization without "de-minning". This figure demonstrates that when the daily load has not been "de-minned" the normalization step causes the signal associated with the relevant variation in electricity usage stemming from the same active consumption behaviors to be significantly muted when there is high baseload and not muted when there is very low baseload. Subsequently moving to the clustering step when this is the case tends to result in one of the resulting representative dictionary cluster having a large membership consisting of undifferentiated flattened load shapes due to the nature of the distance metrics used to score shape fits into best-fit clusters. This means that any information regarding patterns of discretionary electricity consumption behavior is obscured. By “de-minning,” we significantly reduce this problem.[2]

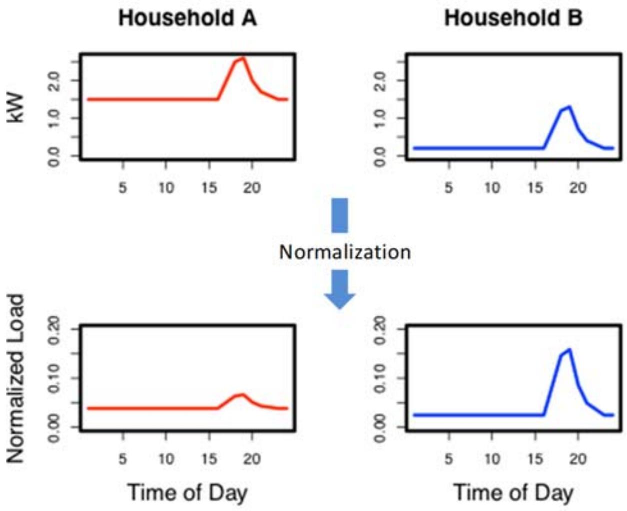

Figure 2 Illustration that normalization without de-minning flattens high baseload profiles.

2.2 Load shape clustering

In the first step of the clustering process we employ, a subset (100,000) of the normalized load shapes first pass through an adaptive K-means (akmeans[3]) algorithm which splits the dataset of load shapes into K_1 clusters, such that the relative squared error (RSE as defined in Equation 1, where s is the load shape of interest, t is the hour of day index, C_i is the dictionary load shape of the cluster that s is assigned to) of any load shape assigned to a cluster is not greater than an error threshold θ. The parameter θ is varied from 0.05 to 0.5 to decide a suitable value that results in the most reasonable K_1.

$$RSE_{s,i} = \frac{\sum_{t=1}^{24}\left(s(t)-C_i(t)\right)^2}{\sum_{t=1}^{24}\left(C_i(t)\right)^2} \tag{1}$$

[2] Specifically, clustering the load profiles into a dictionary of 99 clusters after normalization without “de-minning” resulted in more than 65% of the daily load profiles in the data being assigned to a single flat-shaped cluster. When the daily load profiles were “de-minned” prior to normalization and clustered into the same number of clusters, the highest concentration of load profiles assigned to a single cluster was approximately 10%.

[3] Jungsuk Kwac (2014). akmeans: Adaptive Kmeans algorithm based on threshold. R package version 1.1. https://CRAN.R-project.org/package=akmeans

As the resulting clusters from akmeans are typically highly correlated, in the second step we follow Kwac et al., who propose a subsequent hierarchical merging of the clusters by sequentially combining the most similar clusters until their total count reaches a target number K_2. Under this transformation the guarantee that all RSEs fall under θ is relaxed. The target size K_2 is selected such that it is the smallest number of clusters to violate the θ threshold condition in less than 5% of sample load shapes.

Post clustering processing

In the post clustering processing phase we apply a truncation process to the K_2 cluster centers so that the outlier clusters with low member counts are removed. After truncation, we define the remaining cluster centers to be the final dictionary load shapes. To do this we use an iterative truncation algorithm (detailed in the text box below), so that with violation rate V of user's choice, the maximum number of clusters can be truncated.

```
V: violation rate
Theta: error threshold
while violation < V {
    Identify the smallest clusters whose shape members comprise the
    fraction V of the total number of shapes
    Remove those clusters.
    Reassign the shapes that were members of the removed clusters into the re
maining ones.
    Compute violation rate as fraction of load shapes with RSE > Theta
}
```

Finally, each customer daily load shape is assigned to (and thus represented thereafter by) the closest dictionary load shape.

3 Clustering Results

The resulting number of clusters from the akmeans procedure depends on the chosen error threshold (θ). The relationship derived by running akmeans for θ varying from 0.05 to 0.5 is plotted in Figure 3. We selected $\theta = 0.3$ based on our criteria that the number of clusters compared to the original should not be large (~5K from total of around 30M), and the marginal gain in error improvement to the explanatory power by increasing θ should be small.

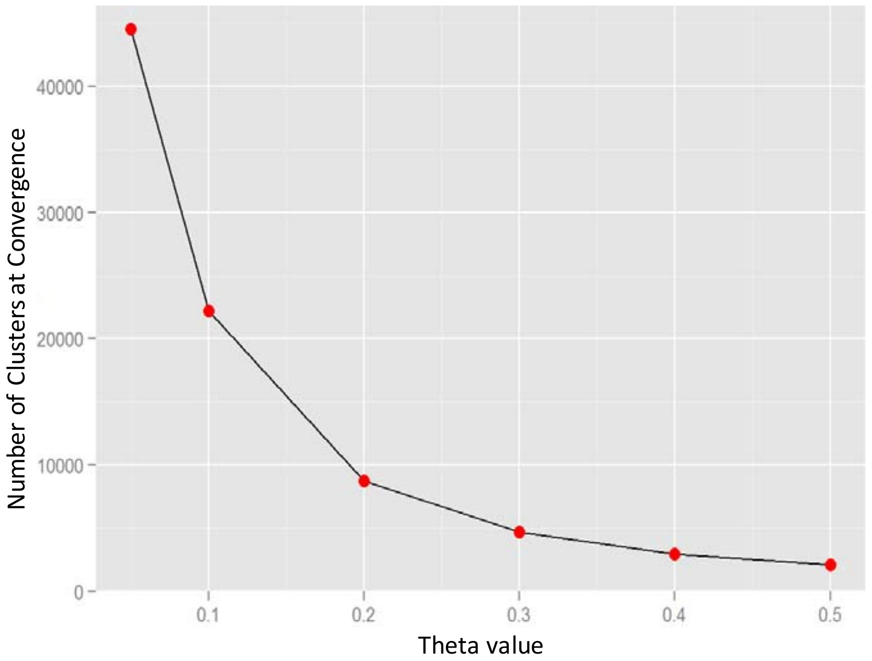

Figure 3 Relationship between error threshold choice and number of clusters from akmeans.

By limiting the total violation rate to 5%, the hierarchical clustering consolidates the number of akmeans clusters to 2000 in our case. The top 600 clusters (sorted by member count) account for ~90% of the data. We use iterative truncation with violation rates of 10% and 30% to further reduce cluster numbers, which results in 608 and 99 clusters respectively. We evaluate these two sets of clusters by using the Davies-Bouldin index (DBI), which measures cluster separation. DBI for the 608 and 99 clusters are 2.23 and 2.22, respectively, indicating similar performance. The 99 cluster centers are then chosen as the final ***dictionary shapes***, and all the load shapes are assigned to their closest centers based on Euclidean distance.

The 16 largest clusters (sorted by membership count) are plotted in Figure *4* with box-whisker plots (i.e. whiskers spanning the 5th and 95th percentiles for each hour of the day). These top 16 shapes account for more than 40% of all the load shapes in the dataset. In this summer-peaking utility, the official time-of-use (TOU) peak rate period is from 4 to 7 PM. While in aggregate, most of the high electricity usage happens in the afternoon, Figure 4 demonstrates that the clusters exhibit considerable differences in peak timing and number of peaks.

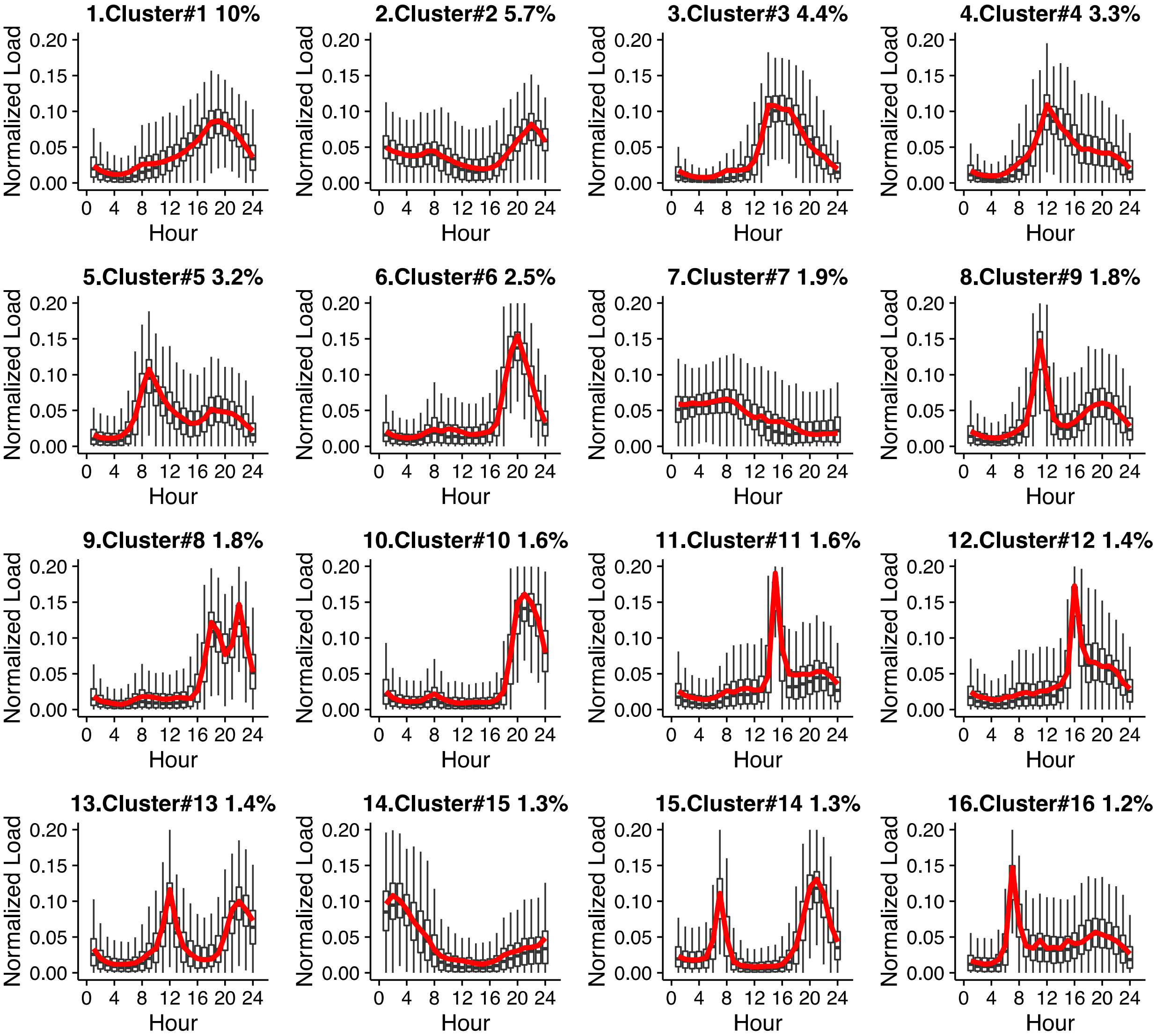

Figure 4 *Top 16 cluster centers. Title format: Rank. Cluster# Percentage of daily loads. Box whiskers summarize within hour distribution of load members belonging to the cluster (median, inter-quartile range, 5th to 95th percentile), with means marked in red.*

The distribution of membership across the clusters in this final 99 dictionary shapes is not uniform, with the largest cluster by member-count accounting for ~10% of the total daily load shapes (left-hand panel of Figure 5). The top 44, 60, and 77 dictionary shapes by member-count respectively cover 70%, 80%, and 90% of the ~30 million load shapes across customers over the entire year period (right-hand panel of Figure 5).

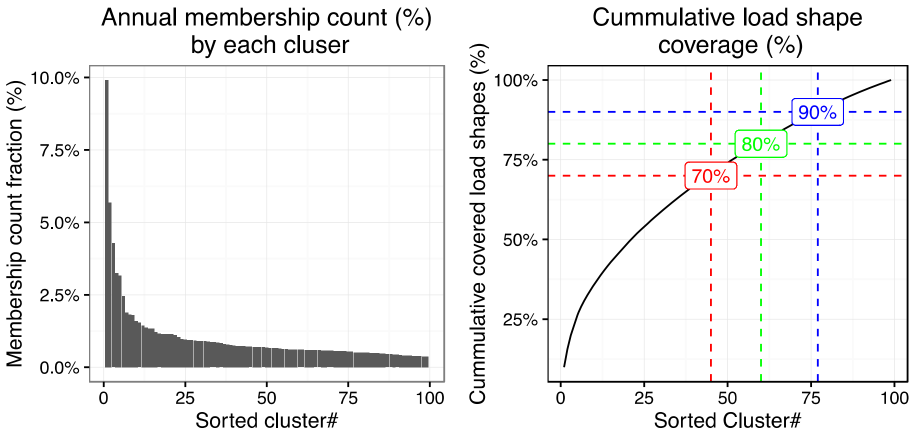

Figure 5 Distribution of load shape membership. Left: load shape coverage by each cluster. Right: cumulative load shape coverage by cluster (sorted by membership counts).

Distribution of electricity consumption is also more concentrated in the top clusters with the largest cluster accounting for ~13% of the total annual electricity consumption (left-hand panel of Figure 6). The top 38, 53, and 73 dictionary shapes respectively cover 70%, 80%, and 90% of annual total electricity consumption of the whole population (right-hand panel of Figure 6).

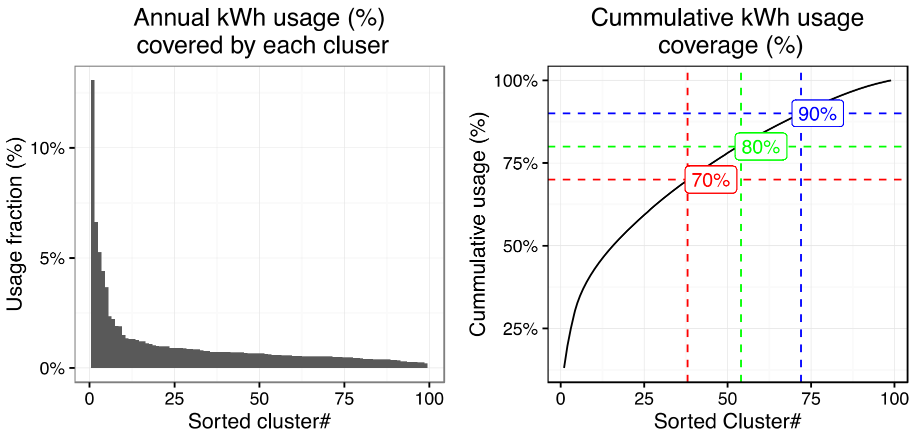

Figure 6 Distribution of energy usage. Left: annual kWh coverage by each cluster. Right: cumulative annual kWh coverage by cluster (sorted by membership counts).

4 Summary and directions for future work

In this technical memo, we employ an innovative clustering technique to categorize daily electricity consumption at hourly resolution from a large sample of residential customers over a full year. “De-minning” is applied to the daily load data (customer-days) in the preprocessing stage so that clustering is focused on the schedules and magnitudes of discretionary consumption. After running adaptive kmeans, performing hierarchical grouping of the resulting clusters, and finally dropping the least populated clusters, a “dictionary” of 99 distinctive “shapes” are identified to represent more than 30 million load shapes within a reasonable error threshold.

Future work using the clustering results will include a demonstration of how consumption patterns can be differentiated by influencing factors such as time scales of interests (seasonal and weekly) and meteorological conditions (outside temperature levels). We will also identify behavioral patterns (i.e. number of usage peaks and timing of major peaks) within the context of a time-of-use rate (i.e., whether households are active during the TOU peak period). Variability within individual households will be characterized by the diversity and composition of the shapes that the respective household possesses.

While typically utilities and system operators focus on the aggregate residential load shape, application of this improved clustering method will shed light on the considerable heterogeneity and variability across days and customers. In the future, more behavioral features associated with household consumption schedules and variability can be extracted based on our results and can be used in future studies to segment customers for better program targeting and designing tailored recruitment strategies.

References

Chicco, G. (2012). "Overview and performance assessment of the clustering methods for electrical load pattern grouping." Energy 42(1): 68-80.

Farhangi, H. (2010). "The Path of the Smart Grid." IEEE Power & Energy Magazine 8(1): 18-28.

Hong, W. C. (2011). "Electric load forecasting by seasonal recurrent SVR (support vector regression) with chaotic artificial bee colony algorithm." Energy 36(9): 5568-5578.

Kwac, J., J. Flora and R. Rajagopal (2014). "Household energy consumption segmentation using hourly data." IEEE Transactions on Smart Grid 5(1): 420-430.

McLoughlin, F., A. Duffy and M. Conlon (2015). "A clustering approach to domestic electricity load profile characterisation using smart metering data." Applied Energy 141: 190-199.

Moslehi, K. and R. Kumar (2010). "A Reliability Perspective of the Smart Grid." IEEE Transactions on Smart Grid 1(1): 57-64.

Rhodes, J. D., W. J. Cole, C. R. Upshaw, T. F. Edgar and M. E. Webber (2014). "Clustering analysis of residential electricity demand profiles." Applied Energy 135: 461-471.

Wang, Y., Q. Chen, et al. (2015). "Load profiling and its application to demand response: a review." Tsinghua Science and Technology 20(2): 117-129.

Zhou, K. L., S. L. Yang and C. Shen (2013). "A review of electric load classification in smart grid environment." Renewable & Sustainable Energy Reviews 24: 103-110.

Zhou, K. L., C. Fu and S. L. Yang (2016). "Big data driven smart energy management: From big data to big insights." Renewable & Sustainable Energy Reviews 56: 215-225.