Creating Representative Load Profiles from Hourly Residential Electricity Load Data - a Clustering Approach

Ricardo Esparza[1], Elisheba Spiller*[1], Karen Tapia-Ahumada[2] and Burcin Unel[3]

[1]*Environmental Defense Fund*

[2]*Massachusetts Institute of Technology Energy Initiative*

[3]*Institute for Policy Integrity at New York University School of Law*

October 12, 2020

Abstract

This paper describes a clustering methodology developed to analyze hourly residential electricity load data to be able to simulate the impact of different rate designs on the adoption of distributed energy resources (DERs). To calibrate the simulation model for DER adoption, we used a dataset of 30 minute electric load intervals for a full year from 44,185 households in Commonwealth Edison (ComEd) service territory in Chicago. Due to computing constraints, we developed a segmentation of the households' data or 'clustering approach' to identify key groups or clusters of households in order to create a smaller number of representative customers of the larger data set.

*Corresponding author: bspiller@edf.org

Contents

1 Introduction

The widespread deployment of Advanced Metering Infrastructure (AMI) has increased the availability of granular electricity consumption data at the household level. AMIs can provide detailed electricity consumption data in short time intervals, but is computationally challenging to deal with such large data sets when analyzing thousands or millions of customers' consumption patterns. There are fluctuations that occur at an hourly, daily, monthly and seasonal level with large variations in usage patterns across customers, and, therefore, it is difficult to concisely capture and describe all of these features. Here, we develop techniques that extract useful and concise information to characterize patterns in these types of large datasets.

Cluster analysis is one of the tools which is frequently used to characterize household energy consumption based on large datasets (Kwac et al. [2014], McLoughlin et al. [2015]). This technique enables researchers and policymakers to better understand and quantify load behavior across a large customer base. Cluster analysis has also been used to improve load forecasting and enhance targeting of energy policies (Yilmaz et al. [2019]).

In this paper, we use a K-means clustering technique to extract segments within a sample of 44,185 households in the Chicago, IL area, using hourly electricity consumption data from Commonwealth Edison in 2016. The objective is to create representative profiles based on these clusters, in order to feed these into an optimization model that simulates the adoption of distributed energy resources (DERs) in response to different electricity prices and rates. Such model is an end-user DER investment and operational engineering optimization model with an adaptation to include an economic utility function, calibrated to the observed hourly residential electric consumption data in order to represent household-level preferences for electricity consumption, simulating the effect of status-quo and more advanced rate design.

2 Data

Our main dataset on electricity consumption is the Five Digit Zip Code Anonymous Advanced Metering Infrastructure (AMI) Interval Data from Commonwealth Edison (ComEd) in 2016, a leap year. The hourly electricity consumption data (in kilowatt-hours, kWh) are collected and created from 30 minute interval data from 44,185 single family residential customers without electric heating for all days in 2016. Table 1 presents the summary statistics of the loads, aggregated to the day, month, and year. These households are located within 3 contiguous zip codes in Chicago. All other characteristics of the household, such as their identity or socioeconomic status, or exact address, are unknown.

Table 1: Distribution of loads across the 44k households in kWh

Variable	1^{st} *percentile*	First Quantile	Mean	Median	Third Quantile	99^{th} *percentile*
Daily Loads	1.5	8.7	19.5	14.7	24.8	78.7
Monthly Loads	57	295	595	480	766	2,128
Yearly Loads	923	4,247	7,151	6,511	9,319	19,138

Figure 1: Mean annual profiles for households in 25, 50 and 75 percentiles of load

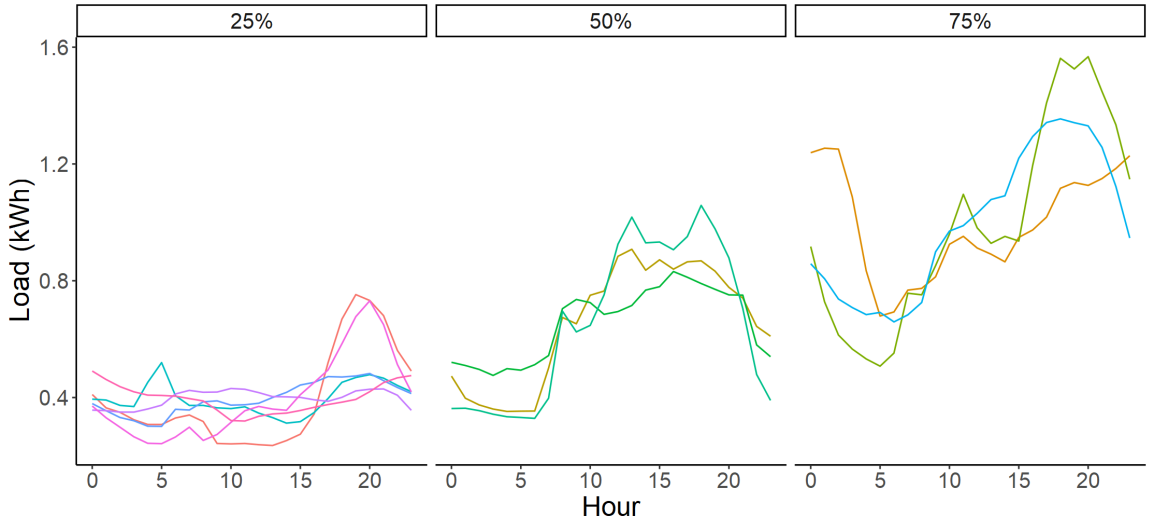
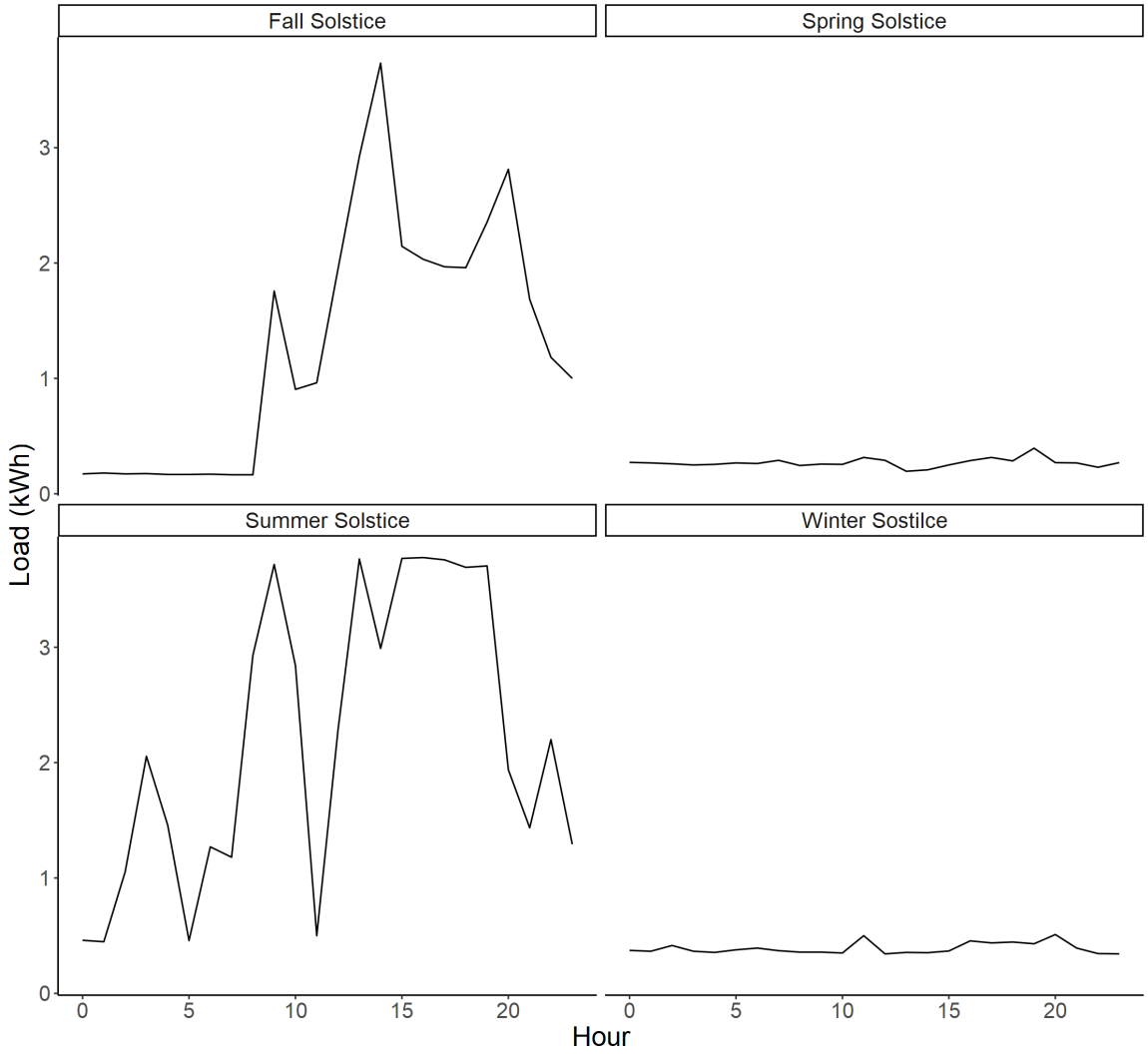


Figure 1 above shows the annual hourly mean load profile for nine different households whose consumption is closest to the first, second and third quartiles of total yearly energy consumption (where each line represents a different household's observed annual average load shape). This figure highlights that even for households with similar total energy consumption, there is a significant variation in load shapes. This phenomenon is exacerbated when load profiles are aggregated to the daily levels, because there also is significant day-to-day variation within each households' annual consumption patterns. Figure 2 below plots one daily profile per season for one specific customer, demonstrating large variations not only across users but even within a household's load profile. Thus, not only is there large intra-household daily variation, but also load profiles typically vary seasonally for each individual household.

Figure 2: Hourly load profile for one customer's daily loads across seasons



3 Methodology

3.1 Characterization of Load Shapes

We cluster based only on electric load characteristics, due in part to data constraints on socioeconomic status, but also because these households are similar in two key ways: they share similar locations (thus, similar weather conditions, power outages, electricity providers, etc.), and they are all single residential households with no electric heating. We

use information on load shape (related to timing of consumption), load size (related to demand), and load variability (related to stability) to cluster, as these are parameters that may impact incentives to invest in DERs.

These three key features of consumption (daily load shape variability, consumption volume, and peak times) are likely to change under different rate designs and thus affect incentives to adopt DERs. Because our purpose is to understand the effect of rate design on DER adoption, we select characteristics of electricity usage within the given data that can account for these three key features.

We create more specific parameters of consumption in the following ways: 1) We define variability to be the variation of daily load shapes across the year; 2) We define consumption volume in terms of daily aggregated consumption and seasonal peak hourly consumption; and 3) We define peak times in terms of the household's peak hours across different seasons. We discuss these in more detail below.

3.1.1 Daily Load Shape Variability

Following Kwac et al. [2014], we express variability in terms of "entropy", which is a way of categorizing the variation in load shapes within a household across the days of the year. We rely on Visdom, a tool in R that was developed by Kwac et al., that is used for organizing and visualizing large datasets. Specifically, we use the package visdomloadshape, which creates a "dictionary" of daily load shapes based on the large dataset of all households' observed loads. The daily load shapes are first normalized by their daily energy demand. The tool then identifies the most common unique load shapes across households using an adaptive K-means unsupervised algorithm (which will be expanded upon further). To simplify the computing process, we use a randomly selected sample of 1,000 households to create the dictionary and a limit of 2,000 different daily load shapes across the entire sample (i.e., 2,000 out of 366,000 possible daily profiles). Finally, with the Visdom-created dictionary of load shapes, we calculate Shannon Entropy per household given its number of different daily load shapes that are encountered within the dictionary. For more details on the process, refer to

Kwac et al. [2014].

3.1.2 Consumption Volume and Peak Demand

We define the consumption volume as the mean daily consumption percentile per household. This was done by creating a distribution of all 366 [1] daily consumption volumes across the whole sample, and then assigning each household into the appropriate percentile for the distribution of all households' daily consumption volume. Then, for each household, we estimate the mean of its 366 different daily consumption volume percentiles.

Another important aspect of household demand used for tariff design and DER investment is the peak demand. Aggregate distribution peak demand determines system capacity requirements and thereby system costs, and may also affect the household's choice of size (or capacity) of the DER technology. Therefore, we also develop another parameter related to the size of each household's mean peak hourly demand. This parameter varies by season, as tariffs are generally seasonally variant, and demand magnitudes are affected by weather conditions. To construct this second demand parameter, we account for both frequency and size, in essence finding the timing of the peak and its magnitude across different seasons. Because there is a household-level variation in both the timing of the peak and its magnitude, we construct the parameter in such a way to reduce the presence and impact of outliers. Thus, to construct the household's average peak consumption, we first identify the household's most common peak hour within the season and then, for that hour, identify the average demand.

3.1.3 Timing of peak demand

Finally, to account for timing of household peak demand, we select the time of occurrence of the aforementioned peak hourly demand. The timing of the peak is critical to identify whether the household's maximum demand occurs at peak time, and whether DERs can provide value and capability to meet this peak.

[1] 2016 was a leap year.

3.1.4 Resulting Characterizing Attributes

In summary, the 6 characterizing attributes per household are:

1. Entropy of encoded daily load shapes' variability
2. Mean daily consumption volume percentile
3. Mean summer peak demand size
4. Mean winter peak demand size
5. Time of summer peak demand
6. Time of winter peak demand

3.2 Clustering the households based on characterizing variables

There are several techniques and approaches to identifying distinct segmented groups within a set of data points: density based clustering, hierarchical clustering, K-means, Gaussian mixture model, and more. Given the nature and scale of our data, we chose K-means as the most appropriate algorithmic technique to cluster the households' electricity consumption. K-means is also a well-known technique (Steinley [2006]), with the capability of being replicated quickly and for larger datasets.

K-means is an unsupervised algorithm that divides N objects (i.e., households), each having observations of P variables (i.e., characterizing variables), into K selected groups or clusters. K-means clusters each household around centroids. These centroids essentially create averages of the characterizing variables across all households within the K groups. The first point is selected at random until the centroid minimizes the Euclidean distance between characteristics of each household and the distances to the centroids of the remaining clusters. K-means is Isotropic, which means that the variables P are assigned equal weight; hence, if variables are not scaled, results could be biased.

Methods for standardization (or scaling of the 6 characterizing attributes) techniques are mixed (Steinley [2006]). However, in our case, standardizing the timing of peak demand

is challenging. Though we input hours into the matrix as numeric data, they are truly a qualitative variable, and even approaching other techniques such as measuring the distance between the peak hour and e.g. 3 pm is not very appropriate. In this case, a hierarchical clustering would be more appropriate to scale all variables; however, given the scale and computing capabilities of K-means, we chose to drop Seasonal Peak Hours from the K-means clustering technique and utilize the peak hours parameters in a second stage of clustering (see below). The remaining variables, excluding Entropy, are already scaled because we are using percentiles. While ideally we would like to scale Entropy, scaling it would lose the power of the concept of Entropy in information theory, which measures how much load shape variability captures each cluster by assigning the probability that each daily load shape repeats in the year. This also allows us to compare different daily load shapes' probability distributions, and normalizing it would change the meaning of the parameter.

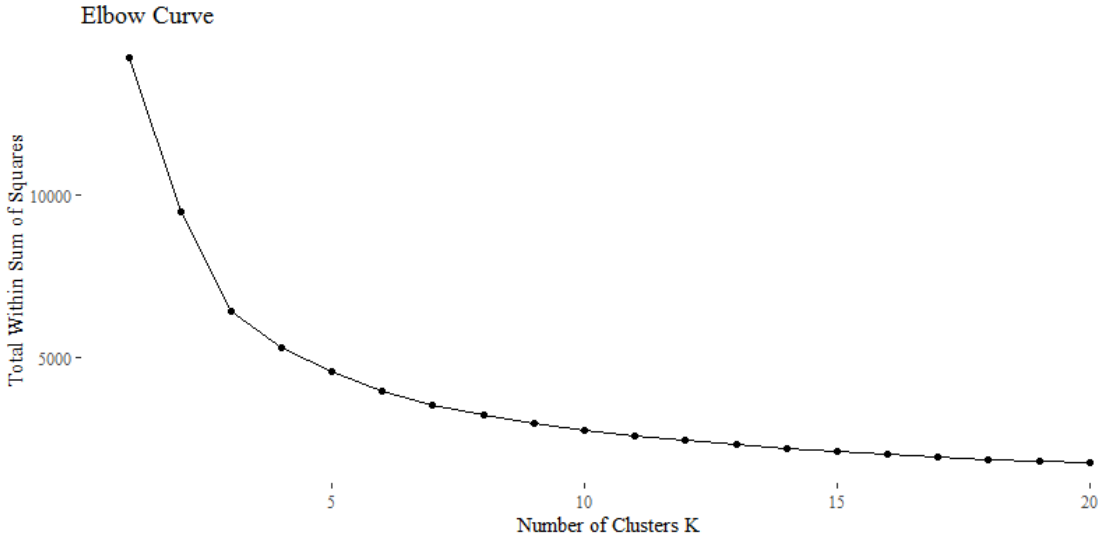
Clusters need to be stable; because the algorithm starts at random, robust clustering should be replicable after many simulations. We employ a bootstrapping technique by Hennig [2007] to measure the stability of our clusters. But the stability relies on the number of K clusters, which is another challenge: k-means works by choosing a number of clusters. The rational for this is that there should be a number of clusters that allows granularity without losing differentiation between each cluster. There are several techniques and heuristics to determine the optimal number of clusters- for example the elbow method (Bholowalia and Kumar [2016]), the gap statistic method (Tibshirani et al. [2001]), the average silhouette method (Rousseeuw [1987]) and the threshold method from the adaptive k-means algorithm (Kwac et al. [2014]).

As a first step, we employ the adaptive k-means algorithm from Kwac et al. [2014], which gives us the optimal number of clusters by iterating from an initially determined number of k clusters and then adding an additional center until no row of characteristics at the household level violates a mean squared error threshold. We chose a threshold of .05, resulting in adaptive k-means finding 5 clusters.[2]

[2] For a more detailed description of the iterative process, see Kwac et al. [2014]

We also verify the number of clusters by using the elbow method, which defines the optimal number of clusters based on where the kink in the curve lies. Figure 3 demonstrates the curve, with the number of clusters on the horizontal axis, and the total sum of squared differences between the observations and their cluster mean on the vertical axis (total within sum of squares).This technique is a commonly-used heuristic. Though 5 clusters is not a very sharp elbow, it does appear to represent a bend in the curve, and thus is not ruled out by this method.

Figure 3: Elbow Method Results



Finally, with the elbow method indicating 5 clusters as appropriate, we verified that the 5 clusters identified by the adaptive k-means technique were stable. According to Hennig [2007], when employing a bootstrapping technique, the score of the procedure (described in more detail in the referenced paper) for each clusters needs to be at least 0.5, and good stability is generally considered to be above .75. The score reflects approximately how many times the cluster was repeated after bootstrapping. After 1,000 bootstraps, the scores for each cluster were:

Table 2: Stability Scores by Cluster

Cluster	1	2	3	4	5
Stability Score	0.70	0.86	0.74	0.85	0.88

The stability scores being all above 0.70 indicate that the 5 clusters found by the adaptive k-means are stable, and that 5 is a good number of clusters according to the elbow method.

Table 3 below shows the characteristics of the 5 preliminary groups or clusters:

Table 3: Summary Statistics of the 5 Initial Clusters

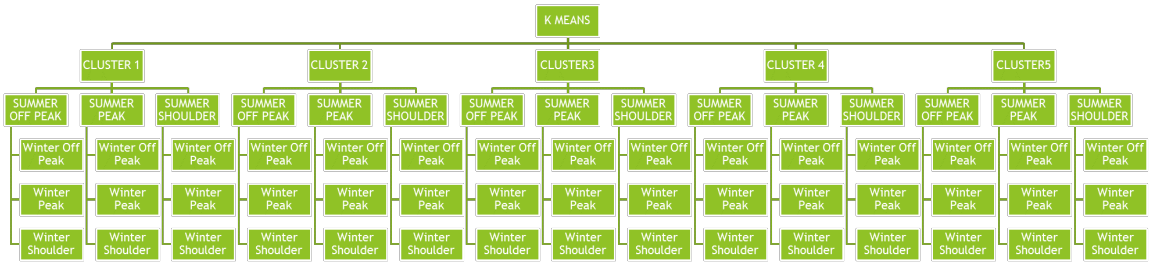
Cluster	Entropy	Daily Percentile	Summer Mean Peak Load	Winter Mean Peak Load	Number of Households
1	7.97	0.61	0.87	0.68	16,770
2	8.01	0.28	0.58	0.31	10,595
3	7.45	0.69	0.87	0.68	8,630
4	6.63	0.50	0.63	0.43	2,516
5	7.48	0.29	0.54	0.27	5,674

As shown here, each cluster represents a variation in the five characteristics. For example, Cluster 1 has high daily load shape variability and high summer peak load, whereas Cluster 4 has low variability and very low winter peak loads.

Now, with an initial set of 5 clusters, we then proceed to include information about the timing of the peak hours. Given the qualitative nature of the peak hour, we separate out the clusters based on different combinations of when the households have their peak demand during the summer and winter. To simplify and to be consistent with the peak hours in the tariffs we are modeling, we divide the hours into 3 periods: Off-Peak (0 to 6), Peak (15 to 19) and Shoulder (the remaining hours). This gives 3x3 possible combinations (depending on what period the customer's peak falls into during the summer or winter), which leads to

a total of 45 different combinations given the preliminary K-means 5 clusters. The decision tree in Figure 4 illustrates this process.

Figure 4: Clustering Second Stage



This results in 45 clusters, and Table 4 below lists the number of households in each cluster.

Table 4: Summary Statistics of the 5 Initial Clusters

Cluster	Combination	Number of Households in Cluster
1	Summer Off-Peak Winter Off-Peak Cluster 1	51
2	Summer Off-Peak Winter Off-Peak Cluster 2	81
3	Summer Off-Peak Winter Off-Peak Cluster 3	42
4	Summer Off-Peak Winter Off-Peak Cluster 4	60
5	Summer Off-Peak Winter Off-Peak Cluster 5	38
6	Summer Off-Peak Winter Peak Cluster 1	95
7	Summer Off-Peak Winter Peak Cluster 2	101
8	Summer Off-Peak Winter Peak Cluster 3	50

9	Summer Off-Peak Winter Peak Cluster 4	63
10	Summer Off-Peak Winter Peak Cluster 5	27
11	Summer Off-Peak Winter Shoulder Cluster 1	320
12	Summer Off-Peak Winter Shoulder Cluster 2	301
13	Summer Off-Peak Winter Shoulder Cluster 3	192
14	Summer Off-Peak Winter Shoulder Cluster 4	167
15	Summer Off-Peak Winter Shoulder Cluster 5	82
16	Summer Peak Winter Off-Peak Cluster 1	243
17	Summer Peak Winter Off-Peak Cluster 2	127
18	Summer Peak Winter Off-Peak Cluster 3	186
19	Summer Peak Winter Off-Peak Cluster 4	101
20	Summer Peak Winter Off-Peak Cluster 5	53
21	Summer Peak Winter Peak Cluster 1	2976
22	Summer Peak Winter Peak Cluster 2	2038
23	Summer Peak Winter Peak Cluster 3	1844
24	Summer Peak Winter Peak Cluster 4	1218
25	Summer Peak Winter Peak Cluster 5	535
26	Summer Peak Winter Shoulder Cluster 1	4655
27	Summer Peak Winter Shoulder Cluster 2	2573
28	Summer Peak Winter Shoulder Cluster 3	2486
29	Summer Peak Winter Shoulder Cluster 4	1528
30	Summer Peak Winter Shoulder Cluster 5	611
31	Summer Shoulder Winter Off-Peak Cluster 1	266
32	Summer Shoulder Winter Off-Peak Cluster 2	181
33	Summer Shoulder Winter Off-Peak Cluster 3	198
34	Summer Shoulder Winter Off-Peak Cluster 4	131

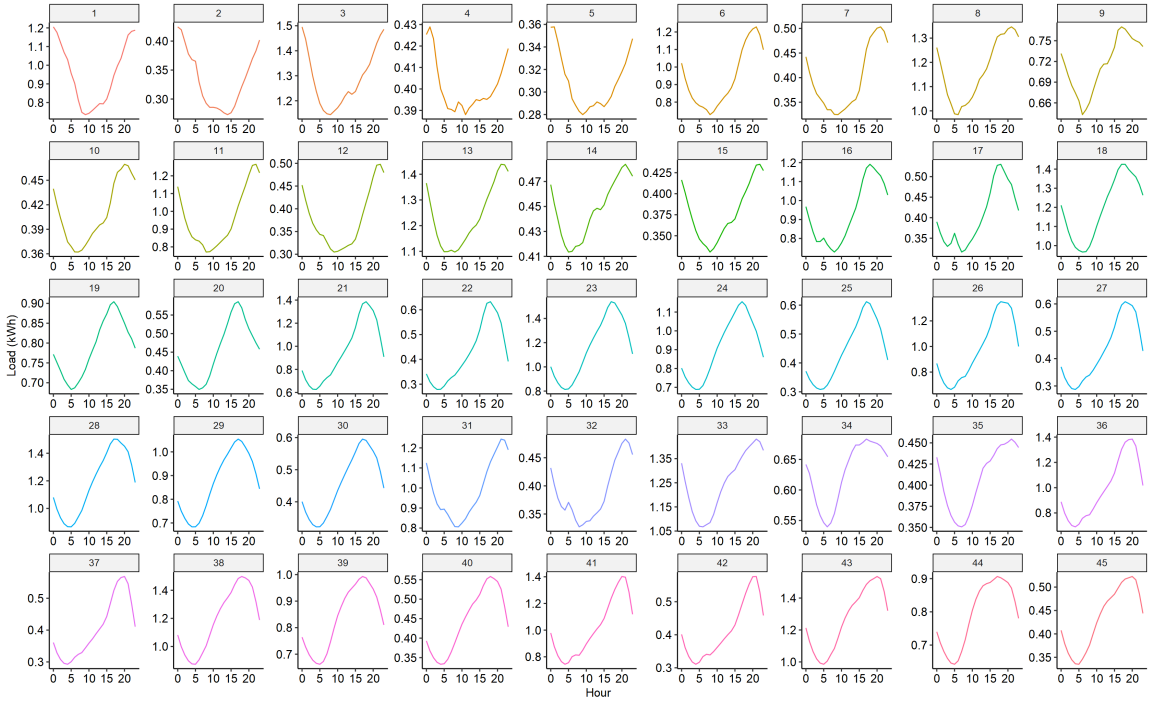
35	Summer Shoulder Winter Off-Peak Cluster 5	86
36	Summer Shoulder Winter Peak Cluster 1	2262
37	Summer Shoulder Winter Peak Cluster 2	1578
38	Summer Shoulder Winter Peak Cluster 3	1180
39	Summer Shoulder Winter Peak Cluster 4	843
40	Summer Shoulder Winter Peak Cluster 5	343
41	Summer Shoulder Winter Shoulder Cluster 1	5902
42	Summer Shoulder Winter Shoulder Cluster 2	3615
43	Summer Shoulder Winter Shoulder Cluster 3	2452
44	Summer Shoulder Winter Shoulder Cluster 4	1563
45	Summer Shoulder Winter Shoulder Cluster 5	741

4 Results

The results of our clustering approach is 45 segmented groups of households, each corresponding to different magnitudes of daily total demand and peak demands, their load stability, and the timing of their peak demand. With this clustering, we can now represent the average load profile per cluster by taking the mean hourly load profile across the year for each cluster.

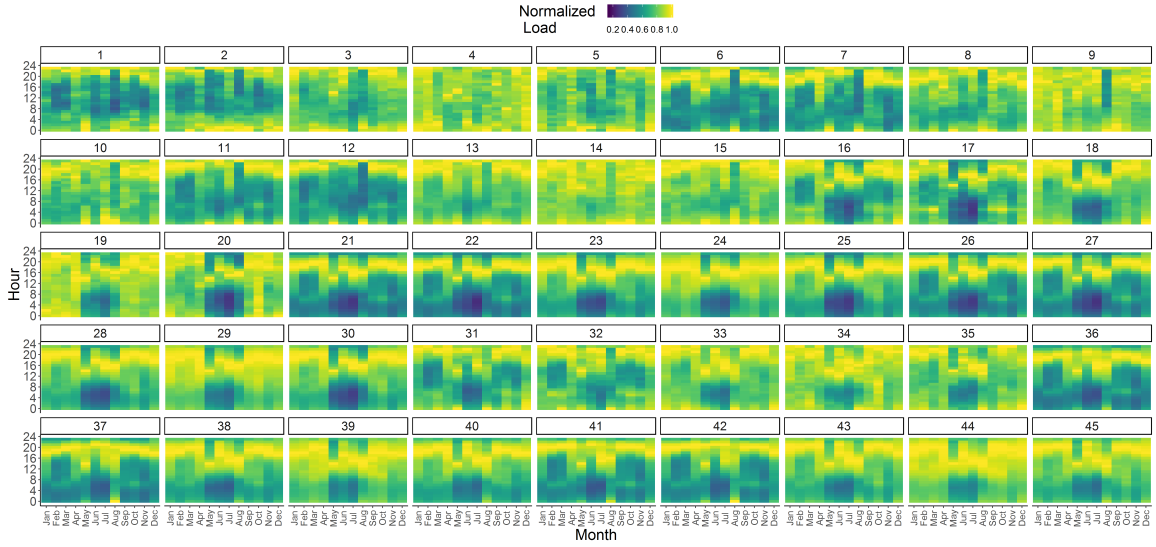
Figure 5 below shows the annual hourly mean profile per cluster. As can be seen in the figure, average load shapes vary based upon magnitudes and timing of peak demands. Thus, the approach was successful in identifying groups of households with different electricity consumption profiles from each other.

Figure 5: Annual hourly mean Load Shapes by Cluster



The heat map in Figure 6 below shows the daily hourly loads normalized by their daily hourly peak load across the years for each cluster. The x-axis represents at a coarser level each day chronologically, with the lighter colors reflecting the hours in which the majority of consumption falls each day. The y-axis represents the hour of each day, hence, consumption should be read as a vertical column that progression in time across the year horizontally. This demonstrates that each cluster, represented in each panel, has a distinct pattern. It also helps to spot seasonality in the households, as well as hourly components. For example, clusters 1 and 2 have high late evening consumption, which explains the shape of their profile. Cluster 42, on the other hand, seems to have stability in its consumption given that the pattern repeats throughout the year, and doesn't show seasonality patterns, which are more visible in other clusters.

Figure 6: Load Heat Maps by Cluster



5 Summary

Our clustering approach summarized an initial sample of 44,185 households to a set of 45 clusters, fulfilling the need to reduce the sample to comply with computing constraints without losing representation of the households' electricity consumption characteristics. Using a mix of quantitative and qualitative techniques, we developed a two-step methodology to create representative load profiles out of the 45 clusters identified.

The approach was successful in finding a statistically sound set of clusters with unique characteristics, with the purpose of saving computing time and power needed. This approach can be used by researchers and policymakers if they wish to characterize large samples of households into smaller groups without having one single representative profile that speaks little of the variability of consumption across the households.

With the increase of more granular electricity consumption data, it is important to find procedures that allow an analysis that is suited to meet time and resource constraints. Clustering is a useful technique to both meet constraints and synthesize the vast amount of information coming from large samples of electricity load profiles. Further research would be

needed to evaluate whether the approach presented in this report would be useful in other applications.

6 Further Research

Because the objective of this clustering technique was to identify groups where the DER adoption incentive and responsiveness to rate design would be unique or similar within each group, and different from the other groups, another way to measure the success of this clustering approach would be to compare those metrics across all clusters (if the data were available). Either way, further work could be done by trying other clustering techniques; for example, testing whether random sampling would have been as effective as the methodology in this report; or a more simple clustering approach such as dividing households by quantiles of energy demand; or constructing clusters based on the quantiles of the number of different daily profiles. Another suggested approach would be to construct a heat map like the one above for all households, and use machine learning techniques to group them according to similarities in those images.

Some other potential next steps could be to further cluster the households- once the main 45 groups have been identified, each main group or cluster could be applied with the same clustering methodology to increase granularity and accuracy of the segmentation. This could lead to a higher number of clusters, but still concise enough to be modelled quickly. It is also important to test whether the attributes chosen for clustering in this report are better than others for segmentation.

Though using household level data without clustering has its advantages, segmenting the data into groups with similar characteristics following the methodology in this report or other methods found in the research literature has the potential to influence policy by identifying the households that are most sensitive to rate design, or whose consumption is highly viable for other interventions like energy efficiency or DER investments. This is also another option to target specific households and induce cost savings/reductions, as well as infer other characteristics in their consumption and inform consumption choices in order to

increase the efficiency of the system.

Clustering provides a tool to increase granularity in synthesis and analysis of high frequency electricity consumption data without analyzing each individual electric customer – which is often prohibitively resource intensive. This method thereby offers a way to move away from the common approach of using a limited set of representative customers and provide utilities, electricity market stakeholders and policymakers with concise but still detailed information from AMI data.

References

Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. 105(9), 2016. ISSN 0975 – 8887.

Christian Hennig. Cluster-wise assessment of cluster stability. 52(1):258–271, 2007. ISSN 0167-9473.

Jungsuk Kwac, June Flora, and Ram Rajagopal. Household energy consumption segmentation using hourly data. 5(1):420–430, 2014. ISSN 1949-3061. doi: 10.1109/TSG.2013.2278477.

Fintan McLoughlin, Aidan Duffy, and Michael Conlon. A clustering approach to domestic electricity load profile characterisation using smart metering data. 141:190–199, 2015. ISSN 0306-2619.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. 20:53–65, 1987. ISSN 03770427.

Douglas Steinley. K-means clustering: A half-century synthesis. 59(1):1–34, 2006. ISSN 2044-8317. doi: 10.1348/000711005X48266.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. 63(2):411–423, 2001. ISSN 1467-9868.

S. Yilmaz, J. Chambers, and M. K. Patel. Comparison of clustering approaches for domestic electricity load profile characterisation - implications for demand side management. 180: 665–677, 2019. ISSN 0360-5442.