K-means Based Clustering Analysis of Household Energy Consumption

Zhenhua Zhang & Timnah Zimet (SUNet IDs: zhenhua & timnah)
Stanford University: CS229A Final Project, 12 December 2018

I. Introduction

The widespread deployment of Advanced Metering Infrastructure (AMI) has made it feasible to investigate into household energy consumption characteristics using smart meter data. Using household energy usage data, we are able to investigate load profiles that encode use behaviors and lifestyles. This information may help tailor different demand response (DR) programs for different households in order to build a more robust grid design system, offer more effective energy reduction recommendations, and improve smart pricing models. Therefore, this project aims at tackling the following questions:

1) How many different clusters characterize representative households?
2) What do load profiles of energy consumption look like for representative households?
3) How many different clusters characterize representative daily usage for a given household?
4) What are the features of these clusters and how should we apply them?

II. Related Work

Various clustering methods have been applied to characterize electricity consumption data, which could be generally grouped into the following categories: hierarchical clustering, partitioning clustering, density-based clustering and model-based clustering (Rokach and Maimon, 2005). For example, Chicco (2012) assessed the performance of different clustering algorithms using 400 electric load patterns during a weekday in Italy with a resolution of 15 minutes, and discovered that whether a cluster algorithm performs well depends on whether the purpose is to identify outliers or to assign all the load patterns to representative clusters. Kim et al. (2011) used 15-min meter data from 3183 customers in South Korea to compare the performance of different clustering methods, including K-means, hierarchical clustering and fuzzy c-means. He found that hierarchical clustering is the optimal approach while K-means is the most efficient option. These studies all adopted normalization to preprocess the data. However, considering we care about the variance as well as the magnitude of the load, we have examined the data both with and without normalization.

Regarding why the K-means method is adopted in this project, an empirical study has suggested that when clustering large data sets, where there are more than millions of records with over 100 dimensions, K-means is the most widely adopted approach because it has linear time complexity and is order-independent (Rokach and Maimon, 2005). Thus, considering that size of the data sets we are plan to deal with, we have chosen the K-means method to conduct clustering analysis of household energy consumption. However, for the final experiment agglomerative clustering is also used, as as comparison study.

Regarding the evaluation metrics of clustering results, there are diverse opinions about which validity indices perform best. Liu et al. (2010) conducted a detailed study of 11 widely used internal clustering validation measures for crisp clustering, and discovered that Calinski-Harabaz Index can handle data set well with subclusters, which are clusters that are close to each other, but cannot give the right optimal number when the data set has skewed distributions or has large noise. Moreover, the Silhouette Index and the Davies-Bouldin Index perform well even with the impact of skwered distributions or noise but cannot handle the problem of subclusters. However, Maulik and Bandyopadhyay (2002) pointed out that Davies-Bouldin Index and Calinski-Harabaz Index perform better than each other when dealing with data sets under different contexts. Therefore, due to the uncertainty of the performance of validity indices, we have used the Calinski-Harabaz Index, Davies-Bouldin Index, and the Silhouette Index to determine the optimal number of clusters.

In terms of the application and interpretation of load profiles generated from electricity consumption data, previous studies tend to focus on charging infrastructure design (Momtazpour et al., 2014) and regression analysis between household features and their load profile clusters (Rhodes et al., 2014). This project expands the perspective of the use of load profile clusters in the electricity market. Pricing schemes and electricity consumption reduction potential have been discussed by interpreting the load profiles; for example, Liu et al. (2018) propose microgrid pricing tariffs based on consumer clusters. As this is a very specific use case and uses normalized data (not accounting for load magnitude), further studies could be beneficial for pricing strategies.

III. Dataset

Data sets containing information about household electricity use and solar panel electricity generation are publicly available on dataport.cloud, a database maintained by Pecan Street research organization. We obtained daily household electricity use data at an hourly resolution over the year of 2016 across 340 households, eventually aggregating a data set of 122,400 records. These households are mostly located in Austin, Texas. The whole data set has been randomly separated into a training set (60%), a validation set (20%) and a testing set (20%).

To preprocess data for the all experiments, the data is reshaped so that each household included in the input is size 365×24, where each row corresponds to a day (one training example) and each hour of the day is a feature. The purpose of this choice of features is to allow us to characterize energy usage by daily use patterns.

In the preprocessing steps for the first experiment (normalized household clustering), we normalize the data by total energy use over each day (see Methods section). This allows us to compare only the shape (i.e. schedule) of daily timeseries data across clusters, as opposed to magnitude of energy usage.

The second and third experiments, the data is used without normalization, since the magnitude of the load is an important factor in grid design and demand response.

IV. Methods

A. Normalizing Data

To normalize data (for some of the experiments), divide the data in each day by total energy use over the day, so that the sum of hourly usage over any one day for any one household is equal to 1. For hourly data point $x_h^{(d)}$ in each day d for each household:

$$x_h^{(d)} = \frac{x_h^{(d)}}{\sum_{h=0}^{23} x_h^{(d)}} \tag{1}$$

B. Clustering Data

K-means clustering was used in all experiments. This clustering algorithm is unsupervised, meaning the training data has no cluster assignment as an input. The algorithm uses the objective function:

$$J = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} - \mu_{c^i}^2 = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_j^{(i)} - \mu_{c^i,j})^2 \tag{2}$$

where $x^{(i)}$ is a training example, and μ_{c^i} is the mean of cluster centroid c that $x^{(i)}$ was assigned to. m is the number of training examples, and n is the number of features in an example.

After randomly initializing the centroids, we optimize J with respect to centroids c^1, c^2, ..., c^K by repeating the following steps until convergence:

- for i = 1:m
  re-assign $x^{(i)}$ to a cluster
- for k = 1:K
  re-calculate cluster centroid μ_k

This is repeated multiple times for multiple random initializations of μ in order to make sure that a global minimum is reached.

In addition to K-means clustering (which we implement ourselves) we use Python's implementation of agglomerative clustering (sklearn.cluster.AgglomerativeClustering) to compare results. Rather than initializing K clusters with random centroids, this algorithm instead begins with every point in the dataset as a "cluster". The two closest points are then combined into one cluster, the new two closest points are then combined into one cluster, and so on. This is repeated until there are only K clusters remaining.

C. Evaluating Metrics

In this section, three clustering validity indices, including Davies-Bouldin Score, Calinski-Harabaz Score and Silhouette Score, that have been used to determine the optimal number of clusters are described in detail as below.

1) Davies-Bouldin Score: This index calculates the ratio of the sum of within-cluster scatter to between-cluster separation. The scatter within the ith cluster, S_i, can be computed as $S_i = \frac{1}{|C_i|}\sum_{x_i}\{||x - z_i||\}$. The distance d_{ij} between cluster C_i and C_j can be calculated as $d_{ij} = ||z_i - z_j||$. Here, z_i represents the ith cluster centroid. Thus, the Davies-Bouldin Score can be defined as:

$$DB = \frac{1}{K}\sum_{i=1}^{K} R_{i,qt}, \tag{3}$$

where $R_{i,qt} = max_{j,j}\frac{S_{i,q}+S_{j,q}}{d_{ij,t}}$.

The objective is to minimize the Davies-Bouldin Score.

2) Calinski-Harabaz Score: The Calinski-Harabaz Score can be defined as:

$$CH = \frac{traceB/(K-1)}{traceW/(n-K)}, \tag{4}$$

where n is the number of data points, K is the number of clusters, B and W are the between and within cluster scatter matrices.

The objective is to maximum the hierarchy level.

Then, the trace of the between cluster scatter matrix B can be computed as:

$$traceB = \sum_{k=1} Kn_k||z_k - z||^2, \tag{5}$$

where n_k is the number of data points in cluster k and z is the centroid of the entire data set.

The trace of the within cluster scatter matrix W can be calculated as:

$$traceW = \sum_{k=1}^{K}\sum_{i=1}^{n_K}||x_i - z_k||^2 \tag{6}$$

Thus, the Calinski-Harabaz Score can be computed as:

$$CH = [\frac{\sum_{k=1}^{K} n_k||z_k - z||^2}{K-1}]/[\frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k}||x_i - z_k||^2}{n-K}] \tag{7}$$

3) Silhouette Score: For each data sample x_i the Silhouette Score can be defined as:

$$Silhouette_i = \frac{d_i - s_i}{max\{d_i, s_i\}}, \tag{8}$$

where s_i is the mean distance to samples in the same cluster, and d_i is the mean distance to samples in the next nearest cluster.

Thus, the Silhouette Score for the whole data set $x_1, ...x_N$ can be computed as:

$$Silhouette = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i - s_i}{max\{d_i, s_i\}} \tag{9}$$

V. Experiments, Results, & Discussion

A. *Normalized Clustering: All Households*

1) Experiment: The first experiment involves clustering all households' normalized data over one year into K use cases, representing different daily schedules of consumers. To determine the optimal number of K, we firstly run clustering methods over the training set for K from 5 to 100 with 1000 iterations, then use three validity indices to evaluate the results. After the optimal K is determined, we run K-means over the training set again with 5000 iterations to obtain the daily schedules.

2) Results & Evaluation: Figure 1 shows the clustering errors of different numbers of clusters based on three indices. For Davies-Bouldin Score, it can be seen that the minimum error is when K is chosen within the range of 5 and 20. For both Calinski-Harabaz Score and Silhouette Score, the optimal K should be chosen as 10 based on the elbow method. Therefore, the final value of K chosen is 10.

3) Discussion: Figure 2 shows the 10 daily representative schedules. In general, different schedules have different peak hours, including 8:00, 13:00, 17:00, 19:00, 21:00, 22:00, with peak loads most commonly occurring in the evening. This could be as a result of the discrepancy of lifestyles, household conditions or other demographic features. Moreover, for each cluster, we can see how much hourly energy use contributed to total energy use.

B. *Non-normalized Clustering: All Households*

1) Experiment: In the second experiment we again cluster all households' data over one year into K use cases, but this time we use non-normalized data. This results in use clusters that account for the magnitude of the load.

2) Results: Again using $K = 10$, household clusters are calculated again using non-normalized data. The centroids are shown in the timeseries plots in Figure 3.

3) Discussion: While most clusters have a peak load in the evening, as in experiment 1, we find that several schedules with very different loads (with peak energy consumption around 0:00-5:00, for example) contribute significantly to total demand. This suggests that most consumers follow normal daytime work schedules but the data may also include industrial use cases, which would be less frequent but have a higher load.

From these 2 experiments we see that both normalized and non-normalized load distributions show important energy use cases. Clustering of normalized loads better shows variance in peak load times, which is useful for peak load reduction planning, while clustering of non-normalized loads may be more useful for targeting different consumers with different pricing or other programs.

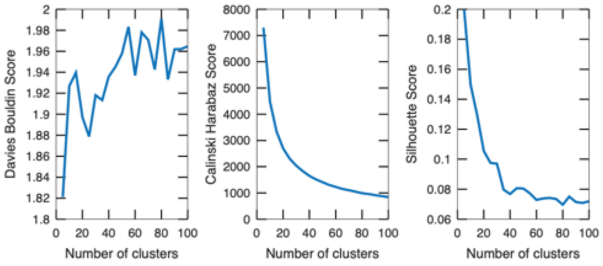

Fig. 1: Validity indices versus the number of clusters

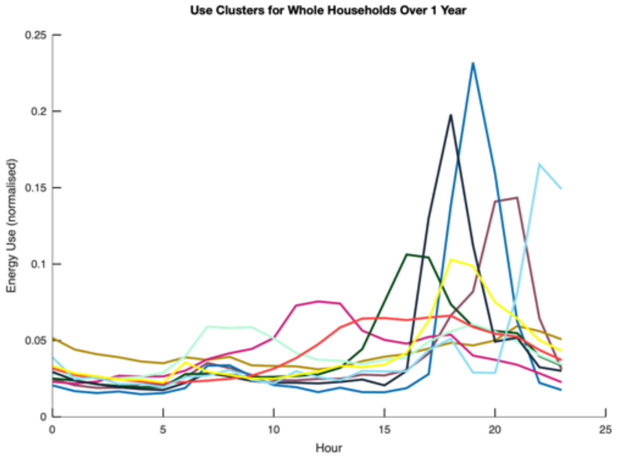

Fig. 2: Representative normalized daily schedules for all households

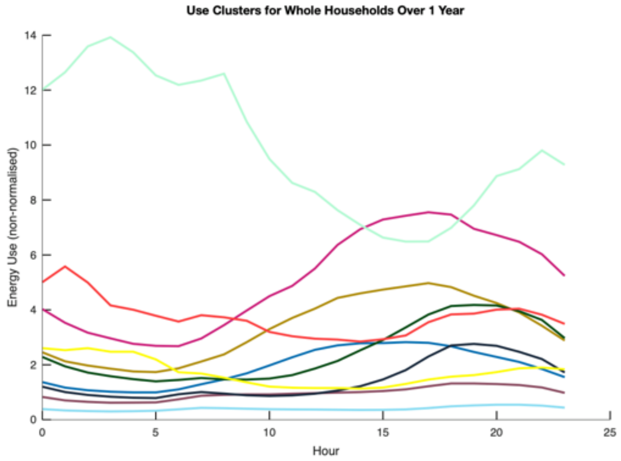

Fig. 3: Representative non-normalized daily schedules for all households

C. *Non-normalized Clustering: Daily Use of Representative Households*

1) Experiment: In this third experiment, for each centroid found in Experiment 1 a single household closest to that centroid is chosen as a representative household. For each of these representative households, 3 non-normalized use clusters are found.

TABLE I: Clustering Composition of Each Representative Household's Use Cases

House-hold	Cluster	# of Days (k-means)	% of Days (k-means)	# of Days (agglom.)	% of Days (agglom.)
8767	1	95	26.0	86	23.5
	2	99	27.0	121	33.1
	3	172	47.0	159	43.4
4767	1	77	21.0	58	15.8
	2	81	22.1	66	18.0
	3	208	56.8	242	66.1
100153	1	85	23.2	21	5.7
	2	122	33.3	129	35.2
	3	159	43.4	216	59.0
8317	1	44	12.0	34	9.3
	2	119	32.5	108	29.5
	3	203	55.5	224	61.2
9939	1	55	15.0	76	20.8
	2	75	20.5	76	20.8
	3	236	64.5	214	58.5
7769	1	69	18.9	80	21.9
	2	93	25.4	103	28.1
	3	204	55.7	183	50.0
5187	1	34	9.3	89	24.3
	2	115	31.4	110	30.1
	3	217	59.3	167	45.6
3482	1	51	13.9	48	13.1
	2	138	37.7	102	27.9
	3	177	48.4	216	59.0
545	1	82	22.4	79	21.6
	2	95	26.0	87	23.8
	3	189	51.6	200	54.6
2575	1	35	9.6	17	4.6
	2	131	35.8	171	46.7
	3	200	54.6	178	48.6

2) Results: Table I shows the composition of the 3 clusters for each representative normalized household. Since the clusters represent typical daily use cases, the table shows the number of days assigned to each cluster, and the percent of days (out of the year of data) assigned to the cluster. The results are shown for both K-means clustering as well as for agglomerative clustering (run using the Python package), for comparison.

The plots of all daily load distributions for each household, colored according to their cluster, are shown in Figure 4 and Figure 5. These show the results for K-means and agglomerative clustering, respectively.

3) Discussion: Since both the representative "schedules" found by clustering normalized data as well as the magnitude of energy consumption are important for grid design and demand response, the results of this section effectively show clusters of clusters that take into account both schedule and magnitude of loads.

The choice of K was based on the expectation that, given a household with a particular normalized use, further variance in magnitude and shape of daily use is due to seasonal changes in irradiation (cluster assignment was found to vary approximately according to time of year), suggesting $K \leq 4$. Visual observation of differences in daily use clusters determines that $K = 3$ captures most of the variance of a household's usage.

We find from the clustering composition evaluation that both K-means and agglomerative clustering perform similarly and skew the clusters a similar amount (i.e. place more data in one cluster than another). The agglomerative clusters are slightly more visually distinct when the features are plotted all at once.

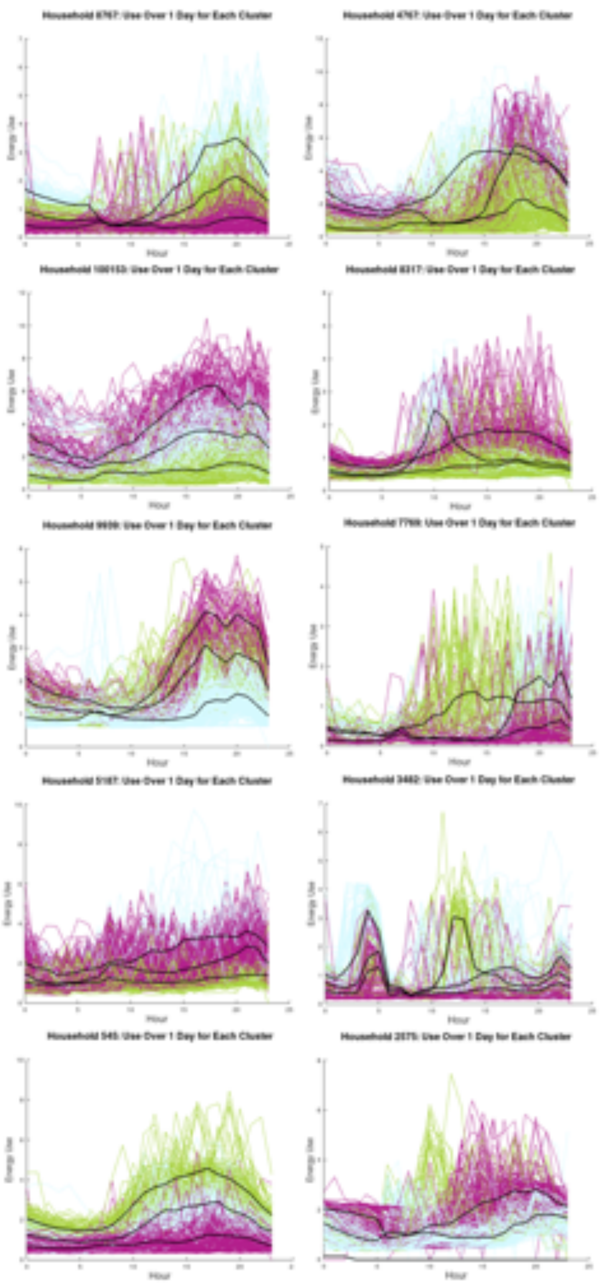

Fig. 4: Representative daily use timeseries clusters for each household closest to a centroid from experiment 1 (K-means)

VI. Conclusions & Future Work

This project expands work on clustering of energy consumption by examining both representative load shapes and magnitudes, as well as representative households and representative 24-hour timeseries consumption. We effectively used unsupervised K-means clustering to cluster households based on 3 evaluation metrics, and further clustered typical use cases of those households at a small value of K. Agglomerative clustering performed similarly; however, further work to compare

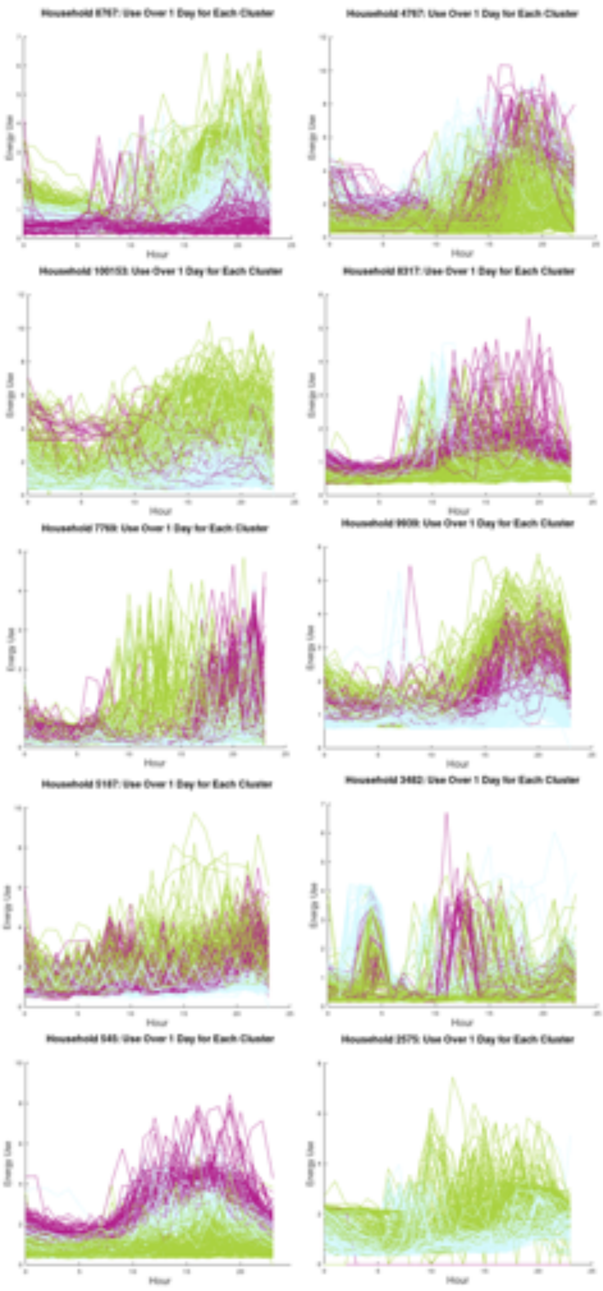

Fig. 5: Representative daily use timeseries clusters for each household closest to a centroid from experiment 1 (agglomerative)

clustering algorithms based on the intended use of the clusters may be useful.

As expected, most load profiles peak in the evening. However, our goal was to group these separately from unusual profile shapes as well as to group different magnitudes/cases of daily usage. This is a key step in targeting specific customer groups with time-of-use pricing to reduce peak load, and may be used in demand response modeling.

As we were not able to obtain data relating the load profiles to local geographical location or socioecenomic profiles, the results of this study are more useful for price modeling and demand response programs than for finding relationships between these demographic factors and energy use. In future work, including features that characterize the households may improve our understanding of the factors associated with different energy use cases, as well as adapting pricing tariffs to take into account low- or high-income housing.

Additional future work could also study methods of combining the results found due to both normalized and non-normalized clustering.

VII. Contributions

Both: contributed to various preprocessing steps and decisions on experiments to try; wrote final report.

Zhenhua: obtained data; wrote and ran K-means algorithm for normalized and non-normalized household clustering; used Python sklearn.metrics packages for evaluation.

Timnah: wrote and ran K-means algorithm for daily timeseries use clustering of K households; used Python sklearn.cluster.AgglomerativeCluster package for comparison; did formatting and layout of poster.

Acknowledgment

Appreciate everything for their existence.

References

G. Chicco. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy*, 42(1):68 – 80, 2012. ISSN 0360-5442. doi: https://doi.org/10.1016/j.energy.2011.12.031. URL http://www.sciencedirect.com/science/article/pii/S0360544211008565. 8th World Energy System Conference, WESC 2010.

Y. Kim, J. Ko, and S. Choi. 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)*, pages 1–6, April 2011. doi: 10.1109/CIASG.2011.5953331.

H. Liu, N. Mahmoudi, and K. Chen. Microgrids real-time pricing based on clustering techniques. *Energies*, 11(6), 2018. ISSN 1996-1073. doi: 10.3390/en11061388. URL http://www.mdpi.com/1996-1073/11/6/1388.

Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916, Dec 2010. doi: 10.1109/ICDM.2010.35.

U. Maulik and S. Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654, Dec 2002. ISSN 0162-8828. doi: 10.1109/TPAMI.2002.1114856.

M. Momtazpour, P. Butler, N. Ramakrishnan, M. S. Hossain, M. C. Bozchalui, and R. Sharma. Charging and storage infrastructure design for electric vehicles. *ACM Trans. Intell. Syst. Technol.*, 5(3):42:1–42:27, Sept. 2014. ISSN

2157-6904. doi: 10.1145/2513567. URL http://doi.acm.org/10.1145/2513567.

J. D. Rhodes, W. J. Cole, C. R. Upshaw, T. F. Edgar, and M. E. Webber. Clustering analysis of residential electricity demand profiles. *Applied Energy*, 135:461 – 471, 2014. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2014.08.111. URL http://www.sciencedirect.com/science/article/pii/S0306261914009349.

L. Rokach and O. Maimon. *Clustering Methods*, pages 321–352. Springer US, Boston, MA, 2005. ISBN 978-0-387-25465-4. doi: 10.1007/0-387-25465-X_15. URL https://doi.org/10.1007/0-387-25465-X_15.