Appliance-level Residential Consumer Segmentation from Smart Meter Data

Lily Buechler and Zonghe Chua
{ebuech,chuazh}@stanford.edu
Department of Mechanical Engineering, Stanford University
Physical Sciences

***Abstract*—The objective of this project was to segment residential consumers based on their appliance-level power consumption from smart meter data for demand response program targeting. Consumers were segmented based on three characteristics of power consumption: their availability (temporal use patterns), variability, and flexibility (willingness to shift power consumption). Four different unsupervised methods were used to segment based on availability from empirical start time distributions of different appliances: K-means clustering, hierarchical clustering, Gaussian mixture models, and latent Dirichlet allocation. Hierarchical clustering most consistently yielded clusters with high load availability during specified hours. Consumer variability was evaluated by clustering load profiles into load profile types and calculating the entropy of the distribution of load profile types assigned to each consumer. Finally, three supervised learning methods were used to predict the responsiveness of consumer power consumption to price based on household characteristics and time-series features. Linear regression with recursive feature selection resulted in the lowest prediction error on the test set.**

***Index Terms*—Consumer segmentation, demand response, unsupervised learning, clustering**

I. Introduction

The increasing deployment of distributed energy resources (e.g. solar, electric vehicles) in power distribution systems will result in greater uncertainty in power demand. One method to mitigate this uncertainty and maintain grid reliability is to enable more control of demand-side resources through demand response programs. Demand response (DR) is a reduction or shift in power consumption relative to baseline behavior during peak loads or high prices. While DR programs have historically focused on the industrial and commercial customers, residential DR programs are expanding. These programs are run by utility companies or third party aggregators and generally focus on control of specific types of residential appliances, such as air conditioners or pool pumps [1].

The effectiveness of a DR program depends on the power consumption patterns of consumers and their responsiveness to prices or incentives. Power consumption at the household level is extremely volatile and can vary significantly from one household to another, given heterogeneity in consumer behavior and the stochastic nature of exogenous variables (e.g. weather patterns). Direct targeting of consumers with behavior patterns well suited for DR would be highly beneficial and cost effective for utility companies.

II. Related Work

Past research has focused on using smart meter data for consumer segmentation to identify households with similar power consumption patterns using unsupervised learning [2]–[6]. An overview of clustering approaches and techniques is provided in [4], and specific relevant studies are highlighted below.

Kwac et al. utilized a combination of adaptive K-means and hierarchical clustering to develop a load profile dictionary from a dataset of 220,000 consumers in CA [2]. Quilumba et al. generated clusters of load profiles using K-means clustering, which they used to improve meter-level load forecasting algorithms [6]. Similarly, Gajowniczek and Zabkowski clustered appliance level activity patterns with hierarchical clustering and used the results to improve load forecasting models [3]. Rhodes et al. analyzed the correlations between seasonal load profiles obtained from K-means clustering and various household characteristics using a probit regression model [5]. Other unsupervised methods that have been tested include support vector clustering, self organizing maps, and fuzzy K-means [6]. All of these studies except for [3] focused on clustering the total load profile of each home, rather than the appliance-level power consumption. However, most DR programs focus on appliance-level control.

III. Project objective

The objective of this project is to segment consumers based on their appliance-level power consumption with respect to three factors: consumer (1) availability, (2) variability, and (3) flexibility. In contrast with previous studies, we focus on appliance-level power consumption. Availability refers to the tendency of a consumer group to consume power during periods of peak system demand when a utility would be most interested in curtailing power consumption. We evaluate consumer availability using unsupervised learning to cluster consumers into groups with similar temporal use patterns for each appliance. Variability is associated with the consistency of power consumption patterns. Consistent use patterns typically result in more accurate power demand forecasts, which improve the effectiveness of a DR program. We evaluate variability by using unsupervised learning to cluster load profiles into discrete groups, and analyze the entropy of the distribution of the load profile assignments for each consumer. Flexibility refers to the willingness of a consumer to shift

TABLE I: Dataset size for each appliance and the type of analysis applied to each (V=variability, A=availability, F=flexibility).

Appliance	Number of homes	Analysis type
Air conditioner (A/C)	129	V
Refrigerator	103	V
Dishwasher	82	V,A
Clothes washer	80	V,A
Dryer	62	V,A
Electric vehicle (EV)	29	V,A
Water heater	13	V,A
Clothes washer + dryer heater	13	V,A
Pool pump	9	V
Total home consumption	132	V,F

TABLE II: Features used for predicting consumer responsiveness to price.

Flexibility features	Units
Home size	ft^2
Year built	year
Number of stories	-
Mean 6-hr energy consumption (24:00-6:00,6:00-12:00,12:00-18:00,18:00-24:00)	kWh
Mean, 10%ile, and 90%ile of daily energy consumption above baseload	kWh
Mean and variance of hourly power consumption	kW
Mean, 10%ile, and 90%ile of maximum and minimum daily energy consumption	kWh
Entropy of load profile (see Section V-B)	-

power consumption from a peak load period to an off-peak period. We use supervised learning to predict the responsiveness of consumers to changes in price based on household characteristics and features extracted from power consumption profiles.

IV. Dataset and Features

Appliance and meter-level real power consumption data was obtained from the open-source Pecan Street Database [7] for homes in Austin TX. Nine of the most common types of appliances were analyzed, and are listed in Table I. Since not all homes had the same appliances, the amount of data obtained for each appliance type varied.

For the availability and variability analyses, twelve months of minute-level data from 2014, 2015, and 2016 were used for the training, validation, and tests sets, respectively. Availability analysis was applied only to deferrable loads, which are appliances that are user-initiated. The power consumption of these appliances is primarily dependent on consumer use patterns. We extracted the start times of each appliance use event from the raw data using thresholding heuristics based on changes in the moving average of power consumption. A multinomial distribution of the start time over the hours of the day was fit for each home and appliance using maximum likelihood estimation with Laplace smoothing. Given an extracted set of appliance start times $S = \{s_1, \ldots, s_N\}$, the probability that an appliance use event for house j occurs during hour h is given by

$$P^{(j)}(s = h) = \frac{\sum_{k=1}^{N} 1\{s_k = h\} + 1}{N + 24}$$

where $P^{(j)}(s) \in \mathsf{R}^{24}$.

We analyzed the variability of all nine appliance types listed in Table I. The 1-minute power consumption data was aggregated to hourly average values, resulting in 365 24-hr profiles for each appliance and consumer. These profiles were then used for consumer segmentation.

For the flexibility analysis, data was obtained from a critical peak pricing study conducted in 2013 on participants in the Pecan Street project [8]. These participants were subjected to a higher electricity price during certain hours (16:00-19:00) on 12 peak pricing days during summer 2013. In the trial, peak days occurred when the predicted maximum daily temperature exceeded a certain threshold. Consumers were notified of the peak pricing event the day before. For this project, we utilized four months of meter-level power consumption data from 32 homes that include all 12 peak pricing days. For each peak pricing day, we calculated the percent reduction in energy consumption during the peak pricing window (16:00-19:00) relative to the mean consumption during the same hours over the previous 14 days. Models were trained to predict the mean percent reduction in consumption over all peak pricing days for each consumer, as a function of various features including household characteristics and attributes of the timeseries power consumption profile, as listed in Table II. Several of these features were chosen based on analysis from [9], which identified specific load profile features that are most indicative of consumer participation rates and response. Daily baseload consumption is defined as the minimum hourly consumption level over the course of a day. The entropy of the load profile of each consumer, which is defined in Section V-B, was also used as a feature. While the availability and variability analysis utilized appliance-level power consumption data, we used the total power consumption of each home to evaluate consumer flexibility. Out of the 32 homes in the dataset, 20 homes were used for training, 6 homes were used for validation, and 6 homes were used for testing.

V. Methods

A. Availability

Four different unsupervised learning methods were utilized for the availability analysis: K-Means clustering, hierarchical clustering, latent Dirichlet allocation (LDA) and Gaussian mixture models (GMM).

For K-means clustering, we ran the algorithm 10 times, each time re-initializing the cluster centers, and chose the cluster assignments with the lowest intra-cluster variation.

For hierarchical clustering, we used a symmetrized version of the KL divergence, which is equal to twice the Jensen-Shannon divergence, as a similarity measure between the start-time probability distributions of consumer i and consumer j for each appliance:

$$\begin{aligned} D(P^{(i)}||P^{(j)}) = &\sum_s P^{(i)}(s) log\left(\frac{P^{(i)}(s)}{P^{(j)}(s)}\right) \\ &+ \sum_s P^{(j)}(s) log\left(\frac{P^{(j)}(s)}{P^{(i)}(s)}\right) \end{aligned} \quad (1)$$

Using Laplace smoothing in the parameter estimation of the probability distributions ensured that $P^{(j)}(s) > 0 \ \forall s = 0, \ldots, 23$ such that Equation 1 is always defined. For hierarchical clustering, we used agglomerative methods with the Ward variance minimization algorithm [10] for the cluster linkage

method. Preliminary results indicated that Ward linkage yields more uniform cluster sizes compared to other methods.

Latent Dirichlet Allocation (LDA) is a generative model typically used in natural language processing that characterizes word "topics" as latent variables [11]. We used LDA since the multinomial distribution of appliance start time probabilities lends itself well to being characterized as a topic modeling type problem. In this application, a "topic" is distribution of start time probabilities and the 24 start times are "words" in a "dictionary". The LDA model was trained by converting the start time probability distributions back into a frequency distribution to use as inputs.

The GMM was modeled as a mixture of 24-dimensional multivariate Gaussians each with tied covariance matrices. This was required since the 24 features are elements of a discrete probability distribution and are thus not independent of each other. The model was trained using expectation maximization.

Both probabilistic models were each run 10 times for 2-14 clusters and the assignments with the lowest intra-cluster variation as measured by symmetrized KL divergence were selected for each of the analyzed cluster sizes.

All four unsupervised learning methods were implemented using the Sci-kit Learn [12] and SciPy [13] Python packages.

Three different metrics were used to evaluate the performance of each method. Suppose the training and validation set both contain n consumers, and k clusters are obtained from both sets such that cluster $c_T \in [0, \ldots, k]$ from the training set contains n_{c_T} consumers and cluster $c_D \in [0, \ldots, k]$ from the validation set contains n_{c_D} consumers. Suppose $c^{(j)}$ is the cluster assignment associated with consumer j. We define the availability of a cluster of appliances during hour h as the mean power consumption during hour h for the entire cluster. The increase in availability from consumer segmentation is the maximum availability over all clusters divided by the availability of all of the appliances in the entire dataset

$$A_\tau = \frac{\max_{c_T} \left[\frac{1}{n_{c_T}|\tau|} \sum_{t \in \tau} \sum_j^n p_t^{(j)} 1\{c^{(j)} = c_T\} \right]}{\frac{1}{n|\tau|} \sum_{t \in \tau} \sum_j^n p_t^{(j)}} \quad (2)$$

where τ is the set of all time indices associated with hour h and $p_t^{(j)}$ is the power consumption of consumer j at time t in the test set. An effective consumer segmentation algorithm should yield a large increase in availability.

The completeness score [14] is a metric generally used to compare a set of cluster assignments to a set of ground truth labels. For this project, we extended the completeness score to the unsupervised case and used it to compare the similarity of clusters obtained from the training and validation sets

$$CS = 1 - \frac{H(C_T|C_D)}{H(C_T)} \quad (3)$$

$$H(C_T|C_D) = -\sum_{c_T=1}^{k} \sum_{c_D=1}^{k} \frac{n_{c_T,c_D}}{n} log\left(\frac{n_{c_T,c_D}}{n_{c_D}}\right) \quad (4)$$

$$H(C_T) = -\sum_{c_T=1}^{k} \frac{n_{c_T}}{n} log\left(\frac{n_{c_T}}{n}\right) \quad (5)$$

where n_{c_T,c_D} is the number of consumers assigned to cluster c_T in the training set and cluster c_D in the validation set. This analysis assumes that consumer power consumption patterns remain similar between the training, validation, and test sets, such that a "perfect" clustering algorithm would recover identical clusters.

Finally, we used the intra-cluster variation of the samples as a measure of cluster quality. The KL-divergence as defined in Equation 1 was used a distance measure between samples. The elbow method [15] was used to select the optimal number of clusters for each appliance.

B. Variability

The variability of consumer power consumption patterns was also analyzed using unsupervised learning. First, K-means clustering was used to cluster the 24-hour power consumption profiles of all homes for each specific appliance into k load shape types. This resulted in 365 load shape cluster assignments for each home and appliance. The distribution over these load shape types $Q^{(j)} \in \mathbb{R}^k$ for each home and appliance over the entire training set was calculated using Laplace smoothing. The entropy of $Q^{(j)}$ gives a measure of the variability of consumer use patterns:

$$S(Q^{(j)}) = -\sum_{i=1}^{365} Q^{(j)}(i) log(Q^{(j)}(i)) \quad (6)$$

C. Flexibility

Three methods were compared for predicting the responsiveness of the power consumption of each consumer to changes in electricity price: linear regression with recursive feature selection, K-nearest neighbors (KNN) regression, and random forests with recursive feature selection. The tuning parameters for each algorithm were selected to minimize the mean squared error (MSE) in the validation set. Tuning parameters included the number of features for linear regression and random forests, the number of neighbors for KNN, and the number of estimators and the maximum tree depth for random forests.

VI. Results and Discussion

A. Availability

An example of the cluster assignments for the start time distributions of clothes washers obtained with K-means clustering for k=5 are shown in Figure 1. The cluster assignments for all four algorithms are qualitatively similar, and effectively segment households by their temporal use patterns. For example, cluster 1 represents consumers with a higher probability of using their clothes washer between 6:00-8:00 and 17:00-19:00. By analyzing such plots, a DR provider would be able to manually select clusters with temporal use patterns most useful for DR program participation. For example, curtailing the power consumption of consumers in cluster 4 could help power system operators meet evening peak system demand.

As shown in Figure 2, availability generally improves as the number of clusters increases. Results show that consumer

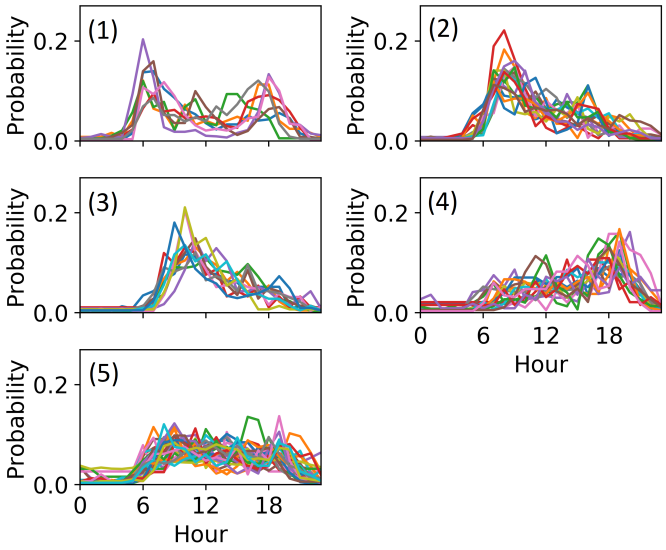

Fig. 1: Cluster assignments for start time distributions of clothes washers for different consumers, obtained with K-means clustering (k=5)

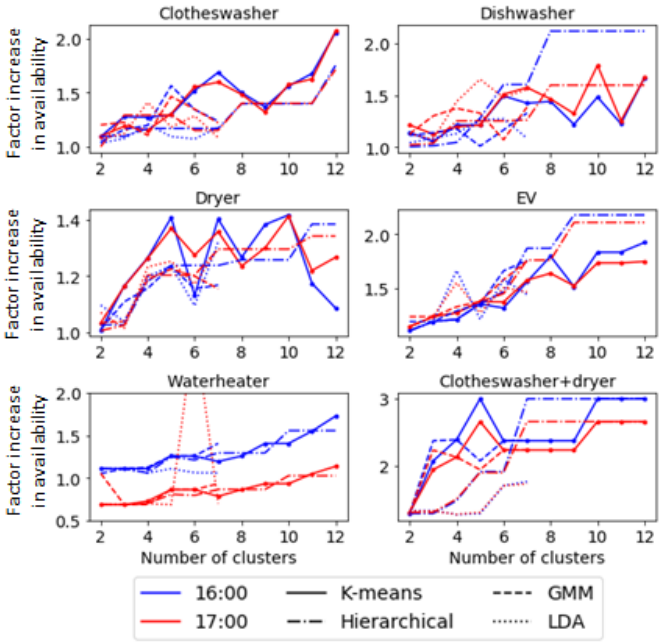

Fig. 2: Increase in load availability from consumer segmentation for different appliances and numbers of clusters

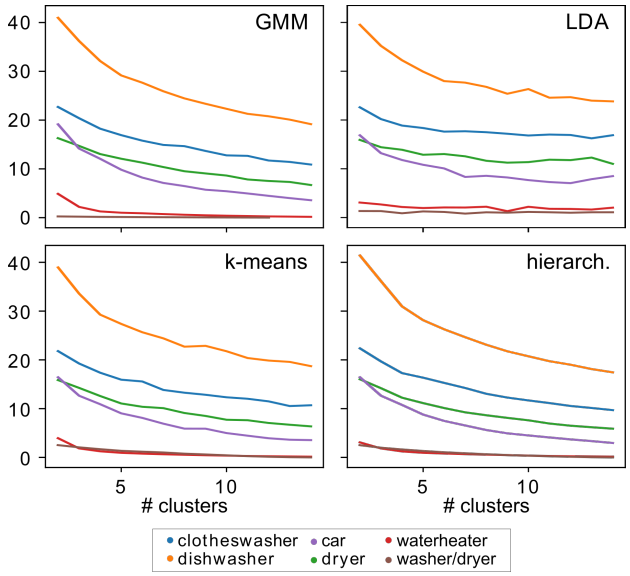

Fig. 3: Intra-cluster variation based on the KL divergence for each unsupervised learning algorithm as a function of the number of clusters

segmentation can result in a x2 increase in the availability of residential loads to participate in DR. While the performance of different algorithms generally depends on the number of clusters, hierarchical clustering consistently results in good performance, especially for large number of clusters. Results also indicate that hierarchical clustering also results in clusters with the most uniform cluster sizes compared with the other algorithms. Additionally, use of KL-divergence as a distance measure between probability distributions is more theoretically justifiable than use of the Euclidean distance in K-means clustering. The performance of the two probabilistic methods may have been limited by the size of the dataset. This is because GMM and LDA use EM-type training methods and thus a dataset where the sample is more typical of the overall population might be more effective. Furthermore, LDA is typically used for natural language processing, where the number of words in the dictionary is significantly larger than the number of hours in a day and the frequency count of words in a corpus is higher than appliance start times in a year.

The use of the elbow method (Figure 3) to find the most appropriate cluster size was difficult due to the fact that there were no clear points where the intra-cluster variation dropped sharply. To aid in our selection of the optimal number of clusters, we also considered silhouette scores, and for LDA, perplexity scores. Though not presented for the sake of brevity, these other evaluation metrics also agreed to a large extent with the clusters found by the elbow method. The number of clusters selected fell into 2 groups: 5 for EVs, dishwashers, clothes washers, and clothes dryers; 3 for washer/dryers and waterheaters.

The completeness scores in Figure 4 were calculated using the selected cluster numbers. Hierarchical clustering had the highest completeness scores for most of the appliances and thus had the most consistent assignments between training and validation sets. The probabilistic methods once again showed poor performance. This was likely due to the same reasons they showed poor performance on the availability metric. Across all algorithms, completeness scores were higher for appliances with smaller number of clusters. This suggests that there is a trade-off between obtaining greater resolution in consumer groups and having higher variance or over-fitting in the clusters.

B. *Variability*

The distribution of the entropy of the load profiles of different appliances are shown in Figure 5 for k=20 (number of load profile types). The mean entropy of the load profiles of deferrable appliances, such as clothes washers, dryers, and dishwashers tend to be higher than that of other appliances.

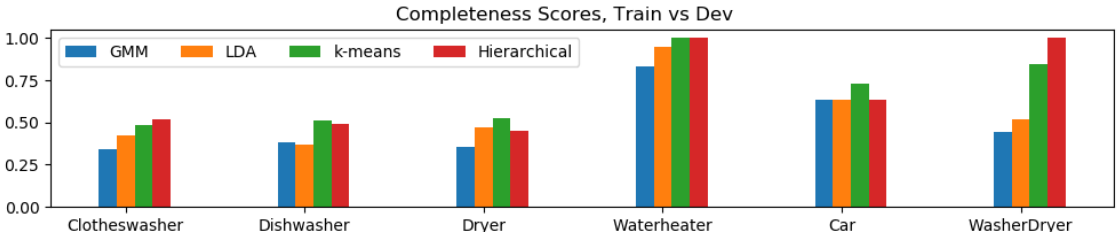

Fig. 4: Completeness score comparison clusters obtained from the training and validation sets for different algorithms and appliance types.

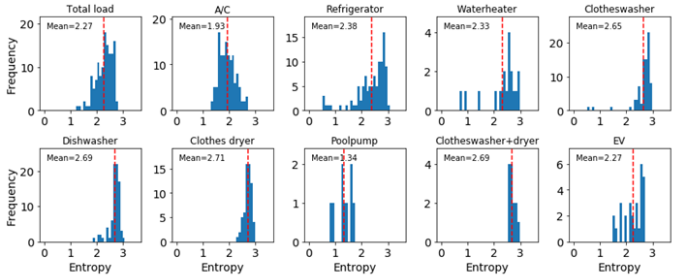

Fig. 5: Histogram of the entropy of load profiles for different consumers for each appliance type

The power consumption profiles of these loads are primarily dependent on occupant behavioral patterns, which can be highly stochastic. In contrast, the mean entropy of the load profiles of thermal loads such as air conditioners, refrigerators, and waterheaters is much lower, but the variance of these distributions is larger. Targeting of consumers with comparatively lower entropy values would be beneficial to utility companies as their behavior is much more consistent on a day-to-day basis.

Increasing the number of load profile types k increases the entropy of the load profiles of all consumers and appliance types. However, it generally does not significantly affect the relative ordering of the mean load profile entropy of different appliance types.

TABLE III: Mean squared errors (MSE) for predicting consumer responsiveness to price.

Method	Train MSE (n=20)	Validation MSE (n=6)	Test MSE (n=6)
Linear regression	0.0200	0.0305	0.0404
K-Nearest Neighbors	0.0148	0.0589	0.0608
Random forests	1.6907	0.7048	0.3187

C. Flexibility

The training, validation, and test mean squared error (MSE) for the three supervised learning algorithms are shown in Table III. The optimal number of features from feature selection was 10 for both linear regression and random forests. For KNN, the optimal number of neighbors was 2 and for random forests the optimal number of estimators was 10 and the maximum depth was 100. Linear regression with recursive feature selection resulted in the lowest test error, followed closely by KNN regression. The three features with highest importance from the recursive feature selection for linear regression were the mean baseload power consumption, the mean energy consumption above the baseload level, and the mean hourly power consumption. Because of the small size of the dataset (n=32), models with lower variance achieved better performance. Random forests would likely have higher performance with a larger dataset.

VII. Conclusions and Future Work

In this project, we developed methods for segmenting residential consumers based on their appliance-level power consumption with respect to their availability, variability, and flexibility. Results indicated that segmentation using unsupervised learning methods can increase load availability for DR programs by up to a factor of two. Hierarchical clustering applied to the start time probability distributions of deferrable appliances produced the most consistent performance. Results indicated that cluster assignments can vary significantly from one appliance to another and can differ from the cluster assignments obtained by only analyzing the total power consumption of each home. This highlights the importance of performing consumer segmentation based on appliance-level power consumption data. Power consumption variability was assessed by calculating the entropy of the distribution of load profile types for individual consumers, identified using K-means clustering. Results indicated notable differences in the variability of power consumption of different appliance types and segments of the population, which could be exploited by a DR program provider. We tested three different supervised learning approaches for predicting consumer responsiveness to electricity prices, and found that low-variance models such as linear regression paired with recursive feature selection resulted in the lowest test error.

Future work may investigate incorporating additional variables, such as day of the week and season into the availability analysis. Expanding the analysis to a larger dataset may provide more insight into the generalizability of the results.

Code for this project can be found at https://github.com/ebuech/cs229.

VIII. Contributions

Lily Buechler performed feature extraction on the raw data, implemented k-means clustering and hierarchical clustering for the availability and variability analysis, implemented the three supervised learning methods for the flexibility analysis, and contributed to the poster and report.

Zonghe Chua implemented the GMM and LDA models and performed the cluster number selection analysis using the elbow method and silhouette scores on all the unsupervised algorithms for the“availability” segmentation. He also contributed to the poster and report.

REFERENCES

[1] Smart Electric Power Alliance (SEPA) and Navigant Research. (2017) 2017 utility demand response market snapshot. [Online]. Available: https://www.navigantresearch.com/reports/market-data-demand-response

[2] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 420–430, 2014.

[3] K. Gajowniczek and T. Zabkowski, "Electricity forecasting on the individual household level enhanced based on activity patterns," *PloS one*, vol. 12, no. 4, p. e0174098, 2017.

[4] Y. Wang, Q. Chen, C. Kang, M. Zhang, K. Wang, and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology*, vol. 20, no. 2, pp. 117–129, 2015.

[5] 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*, vol. 135, pp. 461–471, 2014.

[6] F. L. Quilumba, W.-J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities." *IEEE Trans. Smart Grid*, vol. 6, no. 2, pp. 911–918, 2015.

[7] Pecan Street Inc. (2017) Dataport from pecan street. [Online]. Available: https://dataport.cloud/

[8] N. None, "Technology solutions for wind integration in ercot," Center For The Commercialization Of Electric Technology, Austin, TX (United , Tech. Rep., 2015.

[9] P. Cappers and A. Todd, "Uses for smart meter data, topic 2: Advanced consumer segmentation," 2018. [Online]. Available: https://emp.lbl.gov/webinar/uses-smart-meter-data-advanced-customer

[10] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *arXiv preprint arXiv:1109.2378*, 2011.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed ¡today¿]. [Online]. Available: http://www.scipy.org/

[14] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 2007.

[15] D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: an analysis and critique," *Strategic management journal*, vol. 17, no. 6, pp. 441–458, 1996.