Load Profile Based Electricity Consumer Clustering Using Affinity Propagation

Ahmad Khaled Zarabie, *Student Member, IEEE*, Sahar Lashkarbolooki, *Student Member, IEEE*, Sanjoy Das, Kumarsinh Jhala, *Student Member, IEEE*, Anil Pahwa, *Fellow, IEEE*
Electrical and Computer Engineering, Kansas State University

***Abstract* — With abundant availability of electricity customers load data, and the growing trend toward smart distribution grid, there is a need for more efficient approaches to exploit the valuable customer load information from the high-resolution data collected from customers by automatic meter reading (AMR). New effective clustering methods such as affinity propagation are one of the ways to tackle this issue by improving load prediction techniques and devising efficient pricing schemes. In this paper, an affinity propagation (AP) algorithm is used to cluster customer load data and generate typical load profiles (TLP) for clusters. AP is a new algorithm and has no need to have a predefined number of clusters. Clustering results are compared with some traditional methods such as k-mean, k-medoid, and spectral clustering. Also, the AP results are evaluated by computing a range of well-known clustering performance indices.**

Keywords* — *Affinity Propagation, Clustering, TLP, K-means, K-medoids, Spectral Clustering

I. INTRODUCTION

With the proliferation of smart meters in distribution networks, electricity customers' energy consumption data can be collected vastly more than ever. In a smart distribution system, utility companies and distribution system operator (DSO) can utilize customers' load information to develop and devise proper management techniques which lead to an efficient electricity distribution network. Areas such as load prediction, pricing schemes and demand-side management (DSM) can leverage the abundance of the customer load data. Utilities classify customers into groups based on their energy consumption patterns using their historical load data and generate a TLP considering the time resolution and the high number of customers in a distribution network. The TLPs provide valuable information about a customer to the utility companies and DSOs. In a competitive environment, knowledge of how and when customers use electricity is vital for retailers[1],[2]. TLP can be used for improving the accuracy of the load forecast[3]. Thus, choosing a better clustering algorithm to segment customers is important.

Many clustering algorithms are used for consumer segmentation such as k-means, self-organizing map (SOM), hierarchical and fuzzy. Some of these algorithms are well established and popular like k-means and hierarchical [4]. Clustering algorithms are compared to each other based on their performance, complexity, and operating time. Some perform better than the others in one area but not very good in other areas. Clustering algorithms' quality is determined based on clusters compactness and cohesion which are measured by clustering validation indices. Some important clustering performance indices are Davis-Bouldin indicator, Silhouette, the Dunn index, and many others. They measure the clustering algorithm's performance and behavior.

In [4], Zhou et al provide an adequate review of clustering algorithms and performance indices used in electricity customers segmentation. In [5] authors compare different clustering algorithms with the SOM algorithm in two stages to classify residential customers. Also, they try to draw a correlation between a consumer's load profile and their economic class. However, a two-stage clustering algorithm can be time and resource consuming. In [6] authors use a modified k-mean algorithm to improve the convergence time of the algorithm. They classify consumers into residential, commercial and industrial classes. In [7], a k-mean algorithm is leveraged in two stages to cluster load curves of electricity consumers. First, the authors develop a typical load profile for each customer and then classify customers according to their TLP. Also, performance indices are used to determine the optimal number of clusters. While their work seems promising, using two-stage k-means may not be very practical when the number of customers is high. It may affect the accuracy of the process and be computationally expensive. In [8], the authors use a range of models for load profile, like time domain model, frequency domain model, load shape factor, principal component analysis and so on. They use k-mean and two other modified k-mean algorithms for clustering. Also, various measures are introduced in different papers to validate the classification performance [9], [10], [11].

This paper uses an AP algorithm to cluster residential customer in an electrical distribution network based on their load data. We also implement another three well-known algorithms, k-means, k-medoids and spectral clustering on the same set of data. We compute the six performance validation indices, Silhouette, Calinski-Harabasz, Davis-Bouldin, Dunn index, Ratio of within-cluster sum of squares to between-cluster variation (WCBCR) and cluster dispersion index (CDI) for each algorithm to compare the quality of the clustering. Our objective is to compare the AP algorithm's clustering with other well-known algorithms. AP is a relatively new and simple algorithm. Comparing the clustering quality indices shows that not only affinity propagation outperforms other algorithms, it also addresses some of the drawbacks other algorithms suffer. Partitioned algorithms such as k- means, k-medoids are sensitive to noise, outliers and centroid initialization. Also, the number of cluster must be known beforehand in portioned algorithms[4]. While in AP algorithm, the optimum number of clusters is determined by the algorithm. Unlike k-means and k-medoids, AP is a deterministic similarity-based algorithm.

Main contributions of this paper are listed as followings:

1. This work implements a new clustering algorithm to segment electricity customers based on their load profile
2. This work proposes the AP algorithm that outperforms other major clustering algorithms
3. This work uses the AP algorithm in electricity customers load profile clustering for the first time
4. Results and performance of AP algorithm is compared with three other algorithms

The rest of this paper is organized as follows. Section II introduces the AP algorithm and provides background in k-means, k-medoids and clustering algorithms. Section III presents case studies and results. Section IV provides conclusions and future work.

II. METHODOLOGY

In this work, we use the AP algorithm to cluster *N* residential customers based on their hourly load consumption patterns and generate TLP for each cluster.

AP was introduced in 2007 by Frey and Dueck. It has been used in many clustering applications such as bio-informatics, face and speech recognition[12],[13]. Affinity propagation does not need a predefined number of clusters. It is a deterministic iterative method. AP leverages pairwise similarity between data points and a set of preference measures. Preference measures value show the chances of being a cluster exemplar. In each iteration, a couple of message-passing values are produced. They are called responsibility and availability. Responsibility illustrates how an exemplar is well suited while availability shows how data points are well allocated to a cluster. The detailed procedure is discussed below.

A. Affinity Propagation Algorithm

Hourly customers' historical load data is considered as a data point, $\boldsymbol{x_i}$. The similarity matrix $\boldsymbol{S}$ is devised by computing pairwise distance of points using fifth norm with a negative sign, $\boldsymbol{S}_{i,j} = -\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_{p=5}$. Obviously, the bigger distance between the data points, means that the data points are less similar and vice versa. Number of clusters are not known in as the algorithm starts. The AP algorithm initializes the number of cluster k which are also called exemplars. Exemplars can be initialized in different ways. In this work, AP exemplars k are chosen and initialized by the computing the median of data points' similarity from similarity matrix $\boldsymbol{S}$ as in [12].

Responsibility matrix $\boldsymbol{R}$ and availability matrix $\boldsymbol{A}$ are defined to find the exemplars and cluster the data points. The responsibility value, $\boldsymbol{R}(i,k)$ from responsibility matrix shows how well exemplar k represent point i. At first iteration, responsibility $\boldsymbol{R}(i,k)$ of point i equals its similarity to exemplar k minus its largest similarity to the rest exemplars as in (1). Matrix $\boldsymbol{A}$ equals zero at the first iteration. When $k = i$, the responsibility of diagonal elements $\boldsymbol{R}(k,k)$, is going to be $\boldsymbol{S}(k,k)$ minus the largest similarity between data point i and rest of the exemplars. Negative $\boldsymbol{R}(k,k)$ values mean point k isn't an appropriate exemplar and it has to choose another node as exemplar. So, the responsibility matrix is updated by the following rule.

$$\boldsymbol{R}(i,k) \leftarrow \boldsymbol{S}(i,k) - \max_{k' \neq k}\{\boldsymbol{A}(i,k') + \boldsymbol{S}(i,k')\} \quad (1)$$

In (1), $\boldsymbol{A}(i,k')$, represents the availability of point i with respect to exemplar k'. The points which are allocated to the suited exemplars effectively are going to have a negative availability value.

Availability matrix $\boldsymbol{A}(i,k)$ shows the availability of exemplar k for point i. In other words, availability represents how well it would be for point i to select point k as its exemplar. So, the availability matrix is updated by the following rule.

$$\boldsymbol{A}(i,k) \leftarrow \min\left\{0, \boldsymbol{R}(k,k) + \sum_{i' \notin \{i,k\}} max\{0, \boldsymbol{R}(i',k)\}\right\} \quad (2)$$

To put a threshold on $\boldsymbol{A}\ (k,k)$, the following rule is applied:

$$\boldsymbol{A}(k,k) = \sum_{i'\ s.t.\ i' \neq k} max\{0, \boldsymbol{R}(i',k)\} \quad (3)$$

In each iteration, responsibility and availability matrices are adjusted in (4) to damp their values. Here a constant damping values $\lambda = 0.5$ is used.

$$\boldsymbol{R}_{new\_modified} = (1-\lambda) * \boldsymbol{R}_{new} + \lambda * \boldsymbol{R}_{old} \quad (4.a)$$

$$\boldsymbol{A}_{new\_modified} = (1-\lambda) * \boldsymbol{A}_{new} + \lambda * \boldsymbol{A}_{old} \quad (4.b)$$

The mentioned procedure is repeated until the algorithm converges. The convergence criterion is defined in the following,

$$\boldsymbol{E}^m = \boldsymbol{A}^m + \boldsymbol{R}^m \quad (5)$$

Algorithm's convergence is achieved when $|\boldsymbol{E}^m - \boldsymbol{E}^{m-1}| < \boldsymbol{\varepsilon}$. In (5), m is the number of algorithm iterations. In the end, point k is going to be an exemplar if $\boldsymbol{A}(k,k) + \boldsymbol{R}(k,k)$ is positive. Each point i is being allocated to cluster k based on its maximum similarity to the cluster exemplar.

In this work, we used 5th lp-norm for similarity matrix instead of the original Euclidean distance proposed in [12]. Higher norms give better similarity measures as the higher the norm, the smaller the norm values be. This was verified experimentally while comparing the clustering evaluation indices. Clustering evaluation indices were improved with higher lp-norm.

B. k-means, k-medoids Algorithms

K-means and k-medoids are part of unsupervised clustering algorithms. They are well-known and well-established algorithms. K-means and k-medoids algorithms are explained in details in [14], [15]. Though there are many versions to both algorithms, we only use the traditional algorithms. Both k-means and k-medoids require the number of the cluster to be predefined. In this work, for comparison purpose, the optimal number clusters determined by AP algorithm was used as predefined number of clusters, K for k-means and k-medoids clustering algorithms.

C. Spectral Clustering Algorithm

Spectral Graph Laplacian is another algorithm mostly used for clustering. We use a fully connected graph Laplacian algorithm with Gaussian similarity function proposed in [16], [17] to cluster the electricity customers' load data. Similarity matrix $\boldsymbol{S}$ is computed as below:

$$\boldsymbol{S}(i,j) = \begin{cases} e^{-\left(\frac{\|x_i - x_j\|^2}{\sigma^2}\right)} & i \neq j \\ 0 & i = j \end{cases} \quad (6)$$

In (6), constant σ is the regularizing parameter and x is a data point. Regularizing parameter σ controls the width of the neighborhoods in the graph. The degree matrix $\boldsymbol{D}$ is computed as follows:

$$\boldsymbol{D}(i,j) = \begin{cases} \sum_{i=1}^{N} \boldsymbol{S}(i,j) & i = j \\ 0 & i \neq j \end{cases} \quad (7)$$

The Laplacian matrix $\boldsymbol{L}$ is the difference between the degree and similarity matrices.

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{S} \quad (8)$$

Then the eigenvector and eigenvalues of matrix $\boldsymbol{L}$ are calculated. The two-dimension embedding of the matrix $\boldsymbol{L}$ is computed from the two eigenvectors corresponding to the two smallest eigenvalues. Two-dimension embedded data is segmented by k-means clustering. The same number of clusters K is used as the original k-means algorithm. To choose a suitable value for the regularizing constant σ, the algorithms was run for multiple value of σ. Optimal value of σ was selected by comparing the data point clustering results by computing their quality indices each time.

Once clustering algorithms are implemented, TLPs for each cluster of customers were computed by averaging the load profiles of customers in the same cluster.

D. Clustering Evaluation Indices

To assess the performance of the clustering algorithms, we calculated a range of most-used and well-known clustering evaluation indices including Silhouette, Calinski-Harabasz, Davis-Bouldin, Dunn index, WCBCR, and CDI. These clustering evaluation indices are described thoroughly in [9], [18]-[19]. Detail description of clustering coefficients is not in the scope of this work.

Clustering evaluation indices measure how well data are segmented into classes. Internal clustering quality is often measured by clusters compactness and separation. Compactness shows how well inter-cluster objects are related by computing their variance. Separation measures how much clusters are far apart[11]. For instance, Davis-Bouldin index measures cluster compactness, while Dunn index and CDI measure cluster separation. Indices such as Silhouette, Calinski-Harabasz, and WCBCR indicate to both separation and compactness feature.

III. CASE STUDY & RESULT

In this paper, four clustering algorithms described in pervious section are applied to residential customer's hourly energy consumption data for one year. Customer energy data was accessed from the Pecan Street database[20]. Pecan Street collects customers load data in different time resolutions for over 1000 customer located in different states in the US. One year of customers load data is preprocessed and divided into seasons, weekdays and weekends. The clustering algorithms are implemented on 100 customers for weekdays in the summer season.

The AP algorithm segmented customers into K=7 clusters. Each cluster had different sizes. To make the clustering algorithm comparison valid, we used the same number of clusters K for k-means, k-medoids, and spectral clustering algorithms. Customer clustering result using the AP and k-means clustering algorithms are presented only. We are not reporting the other two algorithms' results to avoid presenting tedious results.

AP clustering results are depicted in Fig. 1. Each plot contains cluster TLP and their corresponding 24-hours customers average load profile. The horizontal axis represents time in hours, while the vertical axis shows power in kWs. The red curves show the TLP of the each cluster and the blue curves represent load profiles for individual customer. Customer clusters have different number of customers, n as shown on subplots. Clusters 5, 6 and 7 have only one customer as they have a unique load pattern. As it can be observed in Fig. 1, customers with similar load characteristic are categorized in the same cluster. This means that these customers have a similar consumption pattern. Though customers in clusters 2 and 3 seemed to have similar profiles, since they have different loading ranges, they are segmented in different clusters. This shows that the clustering algorithms not only classifies customers based on their energy usage temporal patterns, but also considers their consumption ranges. This can be an indication that customers in cluster 3 have a bigger houses or larger families than the customers in cluster 2.

Clustered customers by k-means and their TLPs are shown in Fig. 2. Similar to AP, most customers are classified into four main clusters. Customers with special load patterns are put in separate clusters. Though it is hard to draw a major difference between the algorithms' clustering outputs, there are differences in size of clusters though. For example, cluster 1 by AP in Fig. 1 has more customers than any other clusters by k-means in Fig. 2.

Clustering algorithms' performance are being evaluated by comparing their performance indices. Six clustering performance indices are calculated for each algorithm and shown in Table I. The best values of indices are highlighted in bold text. Though these indices do not have optimal values for best clustering performance, relative enhancement in performance can analyzed by a better performance indices values for the same set of data points. Some indices are better at low values, while some others are better at high values. For instance, smaller values of Silhouette, Davis-Bouldin, WCBCR and CDI indices are better, but larger values of Dunn index, and Calinski-Harabasz indicts to better clustering quality.

Results in Table I shows that the AP algorithm performed better than the other three algorithms based on the measured indices, both in cluster separation and compactness feature of clusters. Four out six indices are better for AP. Specifically, values of clustering performing indices, Silhouette, Davis-Bouldin, WCBCR, and CDI are significantly better for the AP algorithm. The Calinski-Harabasz and Dunn index are better for k-medians and spectral clustering respectively. Among the four algorithms examined, k-means seems to be performing poorly, though its performing indices values are very close to k-median as expected.

AP algorithm has the lowest value of Dunn index. Since it has a better CDI value, one cannot conclude that it performs ill in cluster separation.

Table I: Clustering Performance Indices

Algorithms	Silhouette	Calinski-Harabasz	Davis-Bouldin	Dunn index	WCBCR	CDI
K-means	0.25	23.64	1.60	0.06	0.62	0.16
K- median	0.29	**24.28**	1.34	0.08	0.47	0.11
Spectral ($\sigma^* = 9.5$)	0.33	21.85	1.29	**0.20**	0.40	0.09
Affinity Propagation($\rho^* = -0.43$)	**0.13**	21.19	**0.99**	0.01	**0.34**	**0.08**

σ^* optimal values of regularizing constant, ρ^* initial preference value

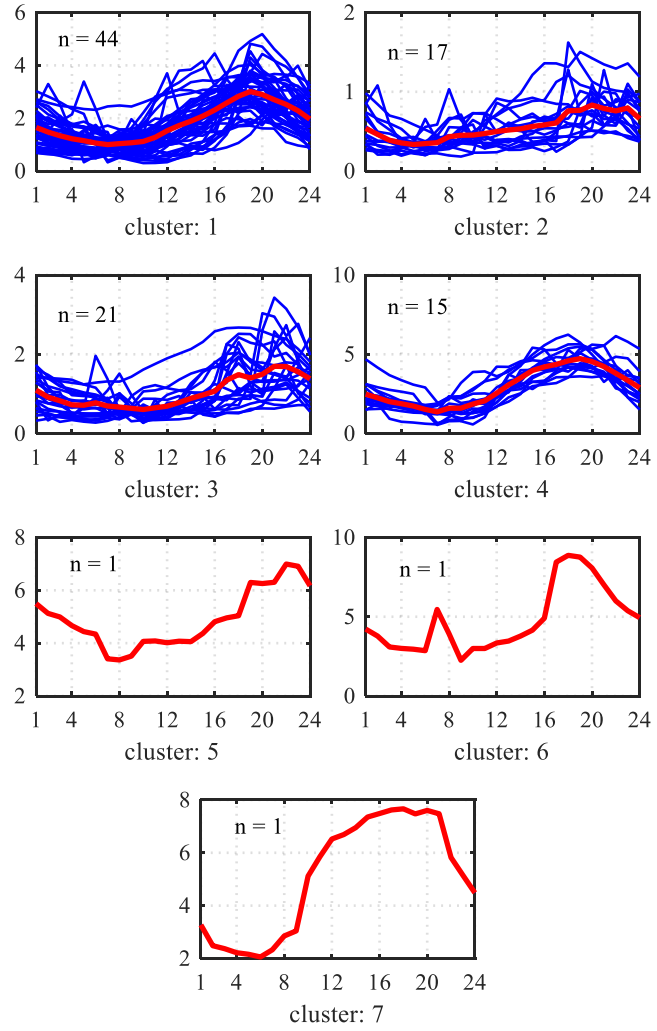

Fig. 1. affinity propagation clustering

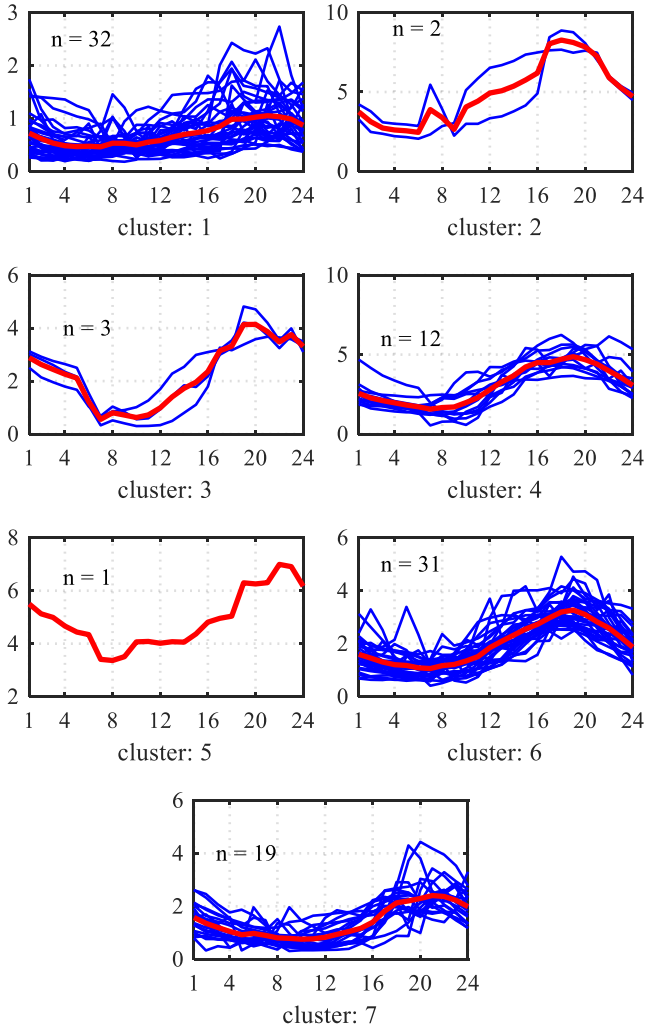

Fig. 2. k-means clustering

IV. CONCLUSION

In this paper, an affinity propagation algorithm is used to cluster customers based on their load data to calculate their TLPs. To illustrate the superiority of the AP algorithm performance, its clustering results are compared to some well-known clustering methods such as k-mean, k-medoids and spectral clustering. The effectiveness of affinity propagation is established through comparing a range of clustering performance indices. Though clustering results from different algorithms are very similar, the computed indices show that the AP algorithm clusters electricity customer better with quality, since 4 out of 6 indices were in its favor. In the future, we want to examine the AP algorithm on a higher number of customers in smaller time resolution and include other electricity customers information such as house size, family size, location, social and economic background, renewable energy generation and weather data This will make the clustering output more accurate which will be important in many application such as load prediction and renewable energy prediction. .

V- REFERENCES

[1] N. Mahmoudi-Kohan, M. Parsa Moghaddam, and M. K. Sheikh-El-Eslami, "An annual framework for clustering-based pricing for an electricity retailer," *Electr. Power Syst. Res.*, vol. 80, no. 9, pp. 1042–1048, 2010.

[2] M. T. Kotouza, A. C. Chrysopoulos, and P. A. Mitkas, "Segmentation of low voltage consumers for designing individualized pricing policies," *Int. Conf. Eur. Energy Mark. EEM*, 2017.

[3] M. Beccali, M. Cellura, V. Lo Brano, and A. Marvuglia, "Short-term prediction of household electricity consumption: Assessing weather sensitivity in a Mediterranean area," *Renew. Sustain. Energy Rev.*, vol. 12, no. 8, pp. 2040–2065, 2008.

[4] K. Zhou, S. Yang, and C. Shen, "A review of electric load classi fi cation in smart grid environment," vol. 24, pp. 103–110, 2013.

[5] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Appl. Energy*, vol. 141, pp. 190–199, 2015.

[6] J. du Toit, R. Davimes, A. Mohamed, K. Patel, and J. M. Nye, "Customer Segmentation Using Unsupervised Learning on Daily Energy Load Profiles," *J. Adv. Inf. Technol.*, vol. 7, no. 2, pp. 69–75, 2016.

[7] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, 2007.

[8] I. P. Panapakidis, M. C. Alexiadis, and G. K. Papagiannis, "Enhancing the clustering process in the category model load profiling," *Transm. Distrib. IET Gener.*, vol. 9, no. 7, pp. 655–665, 2015.

[9] S. Saitta, B. Raphael, and I. F. C. Smith, "A Bounded Index for Cluster Validity," *Mach. Learn. Data Min. Pattern Recognit.*, pp. 174–187.

[10] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, 2002.

[11] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, and S. Wu, "Understanding and enhancement of internal clustering validation measures," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 982–994, 2013.

[12] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science (80-.).*, vol. 315, no. 5814, pp. 972–976, 2007.

[13] X. Zhu, J. Li, Z. Liu, and F. Yang, "A joint grid segmentation based affinity propagation clustering method for big data," *Proc. - 18th IEEE Int. Conf. High Perform. Comput. Commun. 14th IEEE Int. Conf. Smart City 2nd IEEE Int. Conf. Data Sci. Syst. HPCC/SmartCity/DSS 2016*, pp. 1232–1233, 2017.

[14] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3336–3341, 2009.

[15] D. Arthur and S. Vassilvitskii, "K-Means++: the Advantages of Careful Seeding," *Proc. eighteenth Annu. ACM-SIAM Symp. Discret. algorithms*, pp. 1027–1025, 2007.

[16] M. Planck, U. Von Luxburg, and U. Von Luxburg, "A Tutorial on Spectral Clustering A Tutorial on Spectral Clustering," *Stat. Comput.*, vol. 17, no. March, pp. 395–416, 2007.

[17] F. Jordan, "Learning spectral clustering," *Adv. Neural Inf. Process. Syst. 16*, pp. 305–312, 2004.

[18] B. Desgraupes, "Clustering Indices," *CRAN Packag.*, no. April, pp. 1–10, 2013.

[19] S. Saitta, B. Raphael, and I. Smith, "A comprehensive validity index for clustering," *Intell. Data Anal.*, vol. 12, no. 6, pp. 529–548, 2008.

[20] "Electric Consumption Data." [Online]. Available: https://www.pecanstreet.org/. [Accessed: 15-May-2018].