Residential Load Characteristics Analysis Using Clustering and Principal Component Analysis

SON, Ye Ji
Department of Electrical Engineering
Soongsil University
suonj1024@naver.com

AN, Woo Sub
Department of Electrical Engineering
Soongsil University
anwoosub1994@naver.com

BAEK, Sung Jun
Department of Electrical Engineering
Soongsil University
sjbaek@ssu.ac.kr

LIM, Se Heon
Department of Electrical Engineering
Soongsil University
seheon0223@naver.com

YOON, Sung Guk
Department of Electrical Engineering
Soongsil University
sgyoon@ssu.ac.kr

Abstract

Today's electric power systems have been progressed to manage the demand side while the conventional power systems focused on the generation side. With the widespread of measuring device and internet of things (IoT), various types of demand response (DR) programs are being introduced for better utilization of demand resources. To design an appropriate DR program for each consumer, one of the most important steps is to understand the characteristics of each load. In this work, we propose a method to analyze loads' characteristics by using clustering methods and principal component analysis (PCA). The proposed method classifies load data through the k-means clustering algorithm, and then it uses PCA to obtain the principal component of each cluster. To this end, the proposed method uses survey data to analyze the characteristics of consumers. A case study was done with the Irish smart meter data to verify the performance of the proposed method. By using the proposed method, it is expected that utility companies can design a customized DR for each consumer.

Keywords

Residential Load, principal component analysis, k-means clustering algorithm, Demand response

1. INTRODUCTION

Current development of the power system has focused on the demand side management. Thus, the various Demand Response (DR) programs have come out to maximize demand resources. Ratios of total loads to residential loads were accounted for near 22% in U.S. in 2015 and 20% in Ontario, Canada in 2016 [Lopez et al, 2019]. In Korea, it was 13% in 2016 [KEEI, 2018]. It is not a small proportion. Therefore, designing an appropriate DR program for residential consumers is important for utilizing demand resources. It is essential to customize DR design for understanding the residential load characteristics, such as the characteristics of residential consumers and the loads in their houses.

Recent studies have been conducted to utilize the characteristics of residential consumers. In [Haben et al, 2016], Finite Mixture Model-Based Clustering has been proposed for identifying suitable candidates to DR. It helps the distribution network operators manage the Low-Voltage Network. [Chelmis et al, 2015] computed the similarities between consumers with different characteristics, and applied them to DR design. In [Platon et al, 2015], the appropriate characteristics for load forecasting from the various types of buildings have been extracted by using PCA. They used the extracted data as inputs of artificial neural network. [Benitez et al, 2014] conducted a dynamic clustering analysis of residential consumers based on the k-means clustering algorithm. They identified specific consumer patterns for each cluster and analyzed characteristics. In [Vogiatzi et al, 2018], the PCA has been used to identify factors effective on residential buildings for behavior clustering. Many studies use clustering and PCA to analyze the residential load characteristics. Since the accuracy of high-dimensional data clustering is not good, PCA is used to reduce the data dimension in advance as a solution. However, if the clustering is performed after the dimension reduction, the influence of each clusters cannot be identified.

The aim of this study is to suggest the residential load characteristics analysis method using k-means clustering algorithm and PCA. To apply the PCA to individual clusters, we classify the residential consumers through k-means clustering algorithm. Then, PCA extracts the main consumers that can represent each cluster well. Cluster characteristics of extracted consumers are analyzed based on the survey data. This method characterizes the representative pattern characteristics of each existing clusters that can be used for DR design. In addition, residential load characteristics can be estimated without additional survey data. It can be worked with only the classified clusters when a new consumer comes.

The remainder of this paper is organized as follows. In Section 2, we analyze the data used in the case study. Section 3 describes how to characterize the load by using k-means clustering and PCA. Section 4 discusses the results of load characteristics analysis with survey data, and the paper ends with a conclusion in Section 5.

2. DATA ANALYSIS

Case study was conducted on Irish residential smart meter data that was obtained from the Smart Metering Electricity Customer Behavior Trials (CBTs) in Ireland. The 30-minute unit consumption data with more than 5,000 residential and Small and Medium-sized Enterprises (SME) consumers collected from July 14, 2009 to December 31, 2010 [Di cosmo et al, 2014]. After eliminating consumers who did not answer the survey or with missing observations, the number of consumers becomes 3784. Monthly average consumption data for winter (January 2010) are used to consider the monthly

DR design and to find out the relationship with survey question on heating and heat water. The DR events usually happen on weekdays, thus we used only weekday consumption data.

3. LOAD CHARACTERIZATION

3.1 K-means clustering

The k-means clustering algorithm is an unsupervised machine learning method that classifies N sets of multidimensional data $X = [x_1, x_2, \dots, x_N]$ into K clusters. The similarity in the same cluster is increased and between other clusters is reduced so that data having similar patterns are classified.

The algorithm intends to minimize the cost function SSE (sum of squared error). The smaller SSE value means better cluster classification. Where μ_i is the centroid of i^{th}cluster C_i. $\|\cdot\|$ is the Euclidean distance.

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{x_k \in C_i} \|x_k - \mu_i\|^2 \quad (1)$$

The process of k-means algorithm as follow:

Step 1. Set the initial centroids randomly

Step 2. Assign the data to the cluster with the shortest Euclidean distance from the centroid. The following equation is the assignment function. If r_{ki} value is equal to 1, data x_k is belong to i^{th}cluster.

$$r_{ki} = \begin{cases} 1, \text{if } i = \operatorname{argmin}(\|x_k - \mu_i\|^2) \\ 0, \quad \text{otherwise} \end{cases} \quad (2)$$

Step 3. The centroids update to the center of mass in each cluster. Repeat step 2 and 3 until there is no centroid change [Shehroz et al, 2004].

It is important to determine the appropriate number of clusters in advance because k-means clustering is conducted no matter what the optimal number of clusters is. The optimal number was determined through the oldest method called "Elbow method." Increase the number of clusters one by one started from 1, calculate the SSE value of each case, and make graphs [Kodinariya et al, 2013]. If the number of clusters becomes greater than a certain number, a marginal utility of SSE reduction happens and it eventually causes a significant folded distortion in the graph. The folded part is called "elbow" and the number of clusters at the time of elbow occurrence is called "elbow criterion." And, we select "elbow criterion" as an optimal cluster number [Bholowalia et al, 2014].

Fig. 1 shows a graph showing the SSE value of 1≤K≤10. The x-axis is the number of clusters and the y-axis is the SSE value. The amount of SSE reduction is the same as the slope of graph. It is difficult to estimate the optimal number of clusters although the slope can be visually confirmed to decrease at K≥4. Therefore, we calculate the slope difference to figure out the optimal number of clusters. When K changed from 5 to 6, the slope difference was 210, but when changing from 6 to 7, there was a

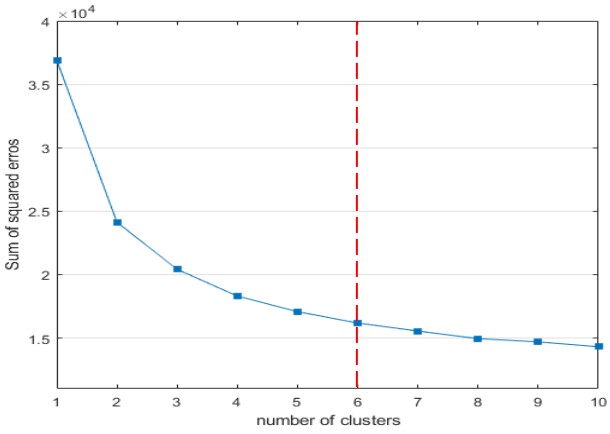

Fig. 1 SSE value (1≤K≤10)

decrease to 63. This shows that the marginal utility occurred at K=6, in other word, the optimal number of clusters is 6.

Fig. 2 shows the consumption of consumers classified as Fig. 2(a) is the monthly average consumption patterns of all consumers, and Fig. 2(b) is the average pattern by cluster. The x-axis and y-axis mean 30-minute unit time and consumption(kWh), respectively. The average patterns in Fig. 2(b) are similar each other, so that the pattern characteristics by cluster cannot be shown. This is because the relatively unimportant consumers are grouped together. Therefore, PCA is used to extract important consumers for more distinct cluster pattern characterization.

3.2 Principal component analysis (PCA)

The principal component analysis method is a technique that reduces the dimension of the existing data $X = [x_1, x_2, \dots, x_N]$ to become more concise by the principal component reflecting the mutual relation between variables. It computes the covariance matrix expressing the relationship of input variables and proceeds singular value decomposition to find a new axis representing the distribution of data [Shlens, 2014].

$$C_x = \frac{1}{N} X X^T \quad (4)$$

The eigenvector of the covariance is called the principal component (PC), and the eigenvector representing the direction with the greatest variance in the distribution of data is the first principal component. An eigenvector having a smaller explanatory variance and being orthogonal to the former principal component is referred to as the second and third principal components [Joliffle, 2011].

The j^{th} principal component PC_j is the linear combination of existing data. The loading vector ϕ_j is the coefficient of the existing variables that constitutes the PC. As ϕ_{jk} value is larger, the variable x_k has the higher importance to the PC_j [Diego Galar et al, 2017].

$$PC_j = \sum_{k=1}^{N} \phi_{jk} x_k \ , \phi_j = \left[\phi_{j1}, \phi_{j2}, \dots, \phi_{jk}\right]^T \quad (5)$$

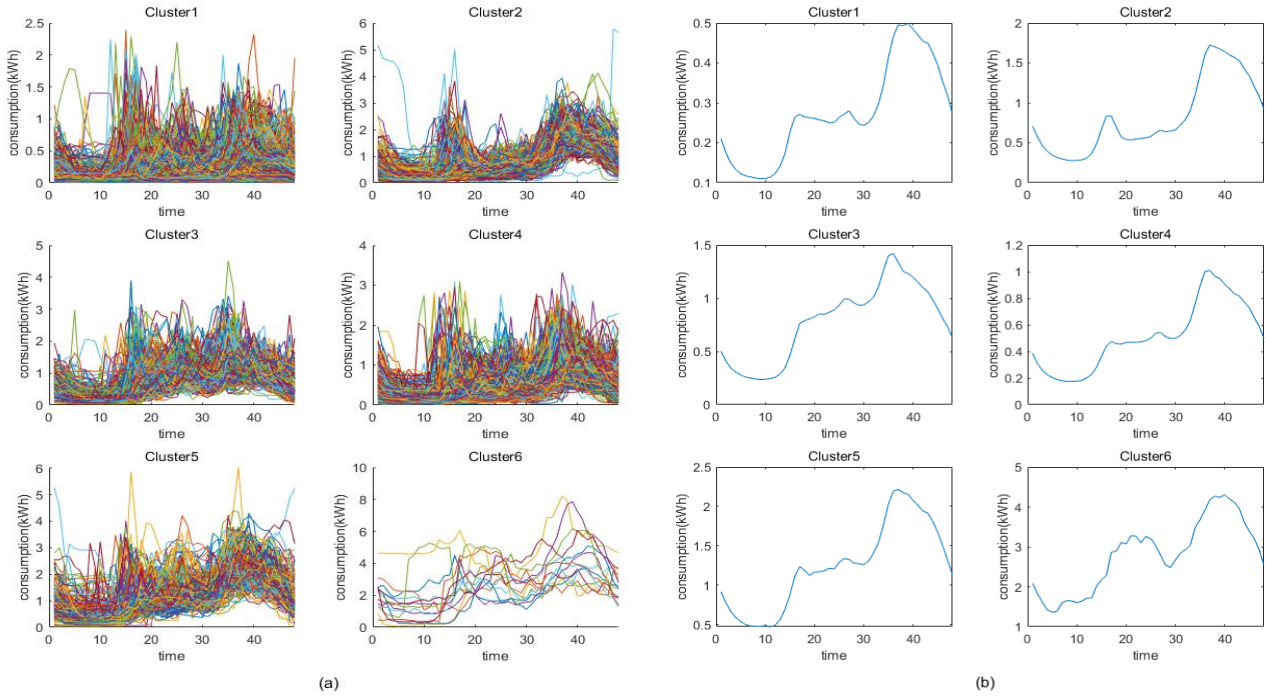

Fig. 2 Consumption pattern by cluster

(a) Monthly average consumption pattern for all consumers (b) Average consumption pattern

EV(j) (Explained variance) is the percentage of the total variance that can be explained by the j^{th} PC. λ_j denotes the variance of the j^{th} PC. The explained variance differs from each PC. And it becomes too small if the number of PC is larger than a certain number. Hence, an appropriate number of PC should be selected. For the reliability of the data to be described as the new principal component, n-principal components should be extracted to account for at least 90% of the overall existing data distribution.

$$EV(j) = \frac{\lambda_j}{\sum_{k=1}^{N} \lambda_k} \tag{5}$$

$$\sum_{j=1}^{n} EV(j) \geq 0.9 \tag{6}$$

For load characterization, it is necessary to remove the consumers who are not important to show the characteristics. The removing criteria is the loading vector. We sort the loading vector $\boldsymbol{\phi}_j = [\phi_{j1}, \phi_{j2}, \dots, \phi_{jk}]^T$ values in descending order and remove the consumers x_k who do not have any ϕ_{jk} within the criteria rank.

$$\text{rank}_i(j) = \lceil (\sum j) * EV_i(j) \rceil$$
$$j = [1,2,\dots,n_i] \tag{7}$$

Ep. 7 is the criteria rank where $\text{rank}_i(j)$ of the j^{th} PC in the i^{th} cluster. n_i denotes the total amount of PC in the i^{th} cluster. The cumulative sum $\sum j$ is the maximum number of the consumer extracted when considering differential importance without EV. Since the EV means the importance of PC, the criteria rank should be calculated differently by EV.

Table 1 shows the number of existing consumers and consumers extracted by PCA. Clusters 6 has 14 consumers.

Cluster	1	2	3	4	5	6
Existing	1021	535	624	1313	277	14
Extracted	75	25	51	57	32	

Table 1 Number of consumers

There is no extraction process in Cluster 6. Because it is possible that only a small number of consumers will be left compared to the others. The number of extracted consumers has decreased to 6.712% of existing consumers.

The average consumption pattern of extracted consumers is shown in Fig. 4. The x-axis and y-axis mean 30-minute unit time and consumption(kWh), respectively.

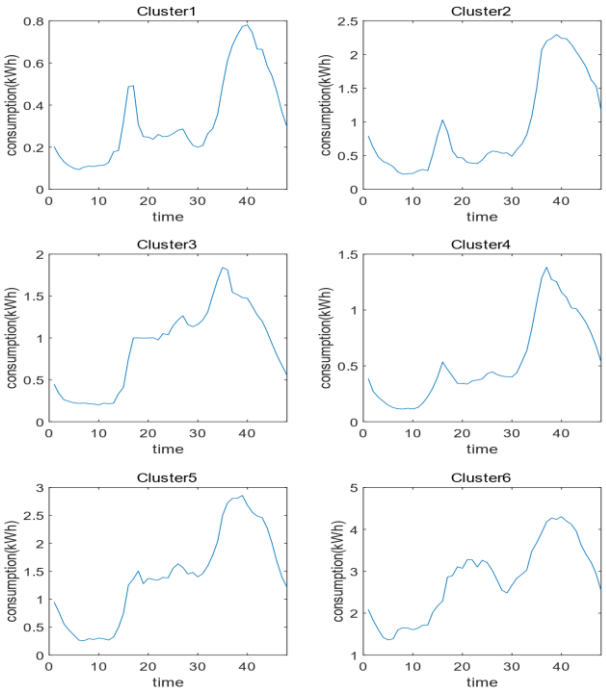

Fig. 3 Average consumption pattern after PCA

Similarity index is needed to confirm that the cluster pattern is well characterized and distinguished from other cluster patterns. The similarity is compared by calculating the Euclidean distance after the min-max normalization of Eq. (8) in Table 2, 3.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (8)$$

$x_{\min}, x_{\max}$ denote the minimum and maximum values of the variable x, respectively. The average distance in Fig. 2(b) was 0.778 and after PCA progression, it increases to 1.205 in Fig.3. Thus, the decrease of similarity between clusters due to pattern characterization can be verified.

Cluster	1	2	3	4	5	6
1	-	0.608	1.007	0.347	0.408	0.781
2	0.608	-	1.474	0.676	0.884	1.335
3	1.007	1.474	-	0.821	0.644	0.826
4	0.347	0.676	0.821	-	0.304	0.863
5	0.408	0.884	0.644	0.304	-	0.688
6	0.781	1.335	0.826	0.863	0.688	-

Table 2 Euclidean distance of Fig.2(b)

Cluster	1	2	3	4	5	6
1	-	0.710	1.719	0.665	1.134	1.432
2	0.710	-	2.059	0.699	1.388	1.746
3	1.719	2.059	-	1.491	0.879	1.080
4	0.665	0.699	1.491	-	1.005	1.429
5	1.134	1.388	0.879	1.005	-	0.651
6	1.432	1.746	1.080	1.429	0.651	-

Table 3 Euclidean distance of Fig.3

4. LOAD CHARACTERSTICS

The characteristics of each cluster were analyzed by using survey from the pre-consumer data collection. After deleting questions that related to the electricity consumption values or with missing answers and duplicate meaning, the number of questions becomes 31 as follows. They are largely sorted into the residential consumers' own characteristics or the load characteristics of their houses.

CONSUMER′S CHARACTERISTIC

1) Family composition
   - I live alone
   - All people in my home are over 15
   - Both adults and children under 15
2) Chief income earner
   - An employee
   - self-employed (w/ or w/o employees)
   - Retired
   - Unemployed (seeking work or not)
   - carer : looking after relative family
3) Number of bedrooms: 1/2/3/4/5++
4) Internet access: Y/N

LOADS′CHARACTERISTIC

1) Heating resource
   - Electricity / gas / oil / solid fuel / renewable
2) Heat water resource
   - Electricity / gas / central heating / oil / solid fuel
3) Number of Appliances: 1/2/3++
   - tumble dryer / dish washer
   - electric heater / stand-alone freezer
   - TV (less than or greater than 21 inch)
   - Desktop / laptop / Games consoles etc.

We calculate the answer ratio per question of each cluster. In "Number of Appliances" analysis, the ratio calculation uses the average number of appliances. Also, the Consumers who use both electricity and other resources were considered when analyzing "Heating and Heat water resource." Fig. 5 shows the survey answer ratio, which is prominently distinguished by clusters among the responses, as a heatmap. The row and column denote each answer option and cluster, respectively. Cluster with the highest answer ratio is displayed in the darkest color.

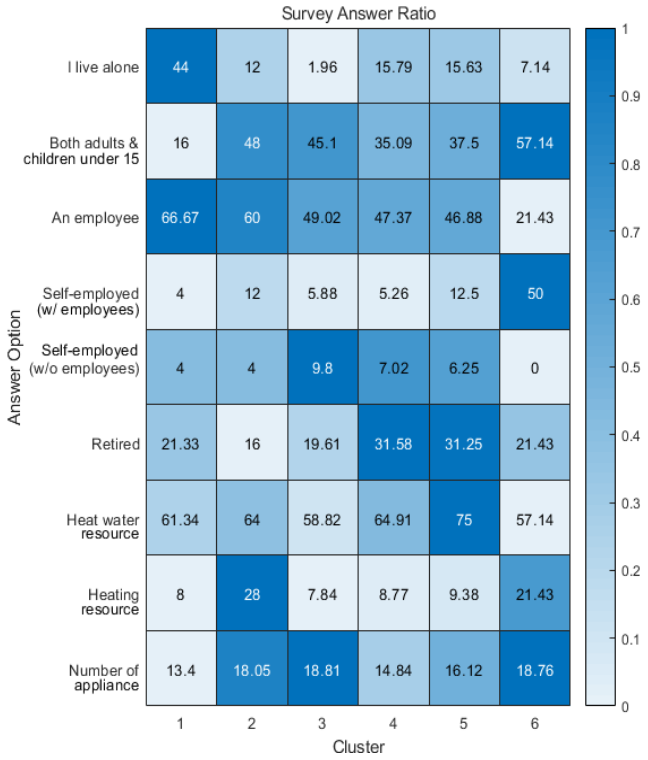

Fig. 4 Heatmap of survey answer ratio

Cluster 1 has the largest proportion of "I live alone" and the lowest proportion of "Both adults and children under 15". Because of a single person, the number of appliances is very small. And the proportion of electricity resource is also low. It is a cause of low consumption in Fig. 4.

Cluster 2 has many appliances and the proportion of electricity resource in "Heating and Heat water" is also large.

The "Number of appliances" proportion of Cluster 3 was very large and most of them are TV, Desktop, and Game

consoles. It makes higher "Internet access" proportion than the others. The answer ratio of "Both adults and children under 15" is similar to Cluster 2 but its power consumption during daytime is clearly larger. It can be assumed that there are preschool children in this cluster.

Clusters 4 and 5 has a large proportion of "Retired". However, the average power consumption pattern of Cluster 5 is higher than the other because the larger proportion of electric resource used in Cluster 5.

Cluster 6 has a low proportion of "An employee" and "self-employed (w/o employees)", while "self-employed (w/ employees)" have a very large share. In Fig. 4, Cluster 6 has a particularly high daytime power consumption because self-employment works at their home. Due to self-employment, "The number of appliances" ratio is very high, and the highest average power consumption was recorded.

We were able to analyze the above characteristics through clustering and PCA. It is possible to deduce characteristics of consumers without survey data according to new consumers are classified into which clusters. The loads' characteristics are the most important part of DR design. When the proportions of "The number of appliances" and electricity resource in "Heating and Heat water" are relatively large, load shifting is being easy. It means that the possibility of participating in DR is higher. So, DR seems to be relatively effective in Clusters 2, 3, 5. Because Cluster 3 has a low share of electricity, Clusters 2 and 5 will have a higher participation rate. In Cluster 6 case, since the power consumption pattern is fixed due to the self-employment. Although there is many shiftable loads in this cluster, the load shifting is difficult. In this case, the energy storage system can be considered.

5. CONCLUSION

In this paper, we proposed a survey-based load characteristics analysis by using k-means clustering algorithm and principal components analysis. The Case study is conducted by using Irish open data. We used monthly average weekday data of winter (2010.01.01 ~ 2010.01.31) after eliminating consumers who did not answer the survey or with missing observations. The elbow method finds out the optimal cluster number. Based on the clustering results, PCA extracts more important consumers to reflect the cluster characteristics. The similarity of the power consumption pattern by clusters is reduced, and the distinction between clusters becomes clearer. After PCA, the load characteristics were analyzed by using survey data. The characteristics of residential consumers and the loads in their house are confirmed. Referring to the load characteristics, we can figure out that the possibility of participating DR of Clusters 2, 3, 5 is high. The customized DR design for each customer becomes possible by using consumer characteristics. When customized time-based DR is designed by using characteristic load pattern and price elasticity of demand, it is possible to calculate consumer's pattern and profits in DR. In future work, we need to suggest an economic analysis of the customized DR program.

Acknowledgements

This work was supported in part by "Human Resources Program in Energy Technology" of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea (No. 20164010201010), and in part by Korea Electric Power Corporation (Grant number: R18XA04).

References

Benitez, I., Quijano, A., Diez, J. L. Delgado, I., 2014, Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers, *International Journal of Electrical Power & Energy Systems*, 55, p437-448.

Bholowalia, P., & Kumar, A., 2014, EBK-means: A clustering technique based on elbow method and k-means in WSN, *International Journal of Computer Applications*, 105(9).

Chelmis, C., Kolte, J., Prasanna, V. K., 2015, Big data analytics for demand response: Clustering over space and time, *2015 IEEE International Conference on Big Data*, p2223-2232.

Di Cosmo, V., Lyons, S., Nolan, A., 2014, Estimating the impact of time-of-use pricing on Irish electricity demand, *The Energy Journal*, p117-136.

Diego Galar, Uday Kumar, 2017, Chapter 5 - Diagnosis, p235-310.

Ding. C., He. X, 2004, K-means clustering via principal component analysis, *In Proceedings of the twenty-first international conference on Machine learning*, p. 29.

Haben, S., Singleton, C., Grindrod, P., 2016, Analysis and clustering of residential customers energy behavioral demand using smart meter data, *IEEE transactions on smart grid*, 7(1), p136-144.

J. M. G. Lopez, E. Pouresmaeil, C. A. Canizares, K. Bhattacharya, A. Mosaddegh and B. V. Solanki, 2019, Smart Residential Load Simulator for Energy Management in Smart Grids, *in IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, p1443-1452

Jolliffe, I., 2011, Principal component analysis, p1094-1096.

Kodinariya, T. M., Makwana, P. R., 2013, Review on determining number of Cluster in K-Means Clustering, *International Journal*, 1(6), p90-95.

Korea Energy Economics Institute, Korea Energy Agency, 2018, Energy Consumption Survey 2017

Platon, R., Dehkordi, V. R., Martel, J., 2015, Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis, *Energy and Buildings*, 92, p10-18.

Shehroz S. Khan, Amir Ahmad, 2004, Cluster center initialization algorithm for K-means clustering, *Pattern Recognition Letters*, Volume 25, Issue 11, p1293-1302.
Shlens, J., 2014, A tutorial on principal component analysis.
Vogiatzi, C., Gemenetzi, G., Massou, L., Poulopoulos, S., Papaefthimiou, S., Zervas, E., 2018, Energy use and saving in residential sector and occupant behavior: A case study in Athens, *Energy and Buildings*, 181, p1-9.