Nigerian Journal of Technology (NIJOTECH)
Vol. 42, No. 2, June, 2023, pp.257 - 263
www.nijotech.com

Print ISSN: 0331-8443
Electronic ISSN: 2467-8821
https://doi.org/10.4314/njt.v42i2.14

LOAD PROFILING OF COMMERCIAL AND RESIDENTIAL BUILDING USING CLUSTERING TECHNIQUE

A. E. Olawumi[1,*], and F. M. Dahunsi[2]

[1]Department of Electrical and Electronics Engineering, Federal University of Technology, Akure, Nigeria.
[2]Department of Computer Engineering, Federal University of Technology, Akure, Nigeria.

*corresponding author (Email: olawumiabayomi@gmail.com, Phone number: +234-806-757-1585)

Article history: Received 13 November, 2021. Accepted 06 July, 2023. Published 31 July, 2023.

Abstract
Data mining is a promising tool used in processing energy data collected from energy consumers. The knowledge derived from energy data is very pertinent in the formulation of various demand-side management programs. This paper uses clustering techniques to segment the energy consumption patterns of residential and commercial buildings; situated at different geographical locations. The two (2) commonly used clustering techniques: K-Means and Agglomerative Hierarchical Clustering, were employed. The result indicates that the choice of clustering technique for load profiling is highly subjective to the nature of the dataset. Hence, using Davies-Bouldin Index (DB) Index and Silhouette Index (SI) as clustering indicators to select an optimum number of clusters and the best clustering technique. Hierarchical clustering was identified as the most appropriate clustering for the two buildings.

Keywords: Data mining, Clustering, Machine learning, Consumption data.

1.0 INTRODUCTION

Analysis of energy consumers' consumption behavior is paramount to the planning and development of the smart grid. In the smart grid, smart meters are used to collect and store energy parameters at different resolutions. The data collected from other geographical locations are transmitted in real-time to a central location where intelligent decisions are taken and reverted to the smart meters [1]. In a modern power grid system, the energy data and other related information are instrumental for applications such as load forecasting, demand-side management, energy theft detection system, and other data-driven operations. Hence, to improve the operations of distribution companies in Nigeria, analyzing consumer's electricity usage patterns is very important in the formulation of tariffs systems and demand-side managements that can minimize energy usage at peak hours.

Due to the 4V (volume, variety, velocity, and value), challenges identified with smart meter data [2], a more intelligent approach is needed to analyze energy data for any application. Data Mining (DM) has been identified as the best computational technique. Among the various DM techniques used on smart meter data, classification and clustering are the most used. Clustering techniques intelligently classify common patterns into the same group, which is more effective than statistical methods as it prevents loss of energy estimates [1] [3].

In this paper, clustering techniques will be used to analyze the consumption patterns of energy consumers. In Nigeria, residential and commercial sectors represent about 80% of electricity demand [4]. However, among the various clustering techniques, K-Means and Hierarchical Clustering are well known for high performance [5] [6]. K-Means have been identified as the most efficient algorithm among partitional clustering techniques. Also, Hierarchical Clustering is known to require less processing time while clustering [7]. To this end, Agglomerative Hierarchical Clustering and K-Means were used to cluster the consumption behaviors of a residential building and a commercial building.

2.0 BASIC METHODOLOGY

Clustering is an unsupervised learning algorithm that intelligently classifies objects with similar attributes in a group. The algorithm is capable of bringing out hidden similarities from sparse data. It is widely employed in various disciplines such as image processing, pattern recognition, etc. This work used the technique to group the different data points acquired from the normalized data from 'OSU' and 'OND'. The method of determining the hidden similarities varies, hence, the existence of disparate clustering techniques. Partitional and Hierarchical Clustering algorithms are the most ubiquitous and have been identified to be efficient [7]. The partitional clustering algorithm identifies patterns in a dataset by optimizing a specific objective function and iteratively improving the quality of the partitions. Hierarchical Clustering group data objects by developing a binary tree-based data structure called the dendrogram. After the dendrogram has been built, the right number of clusters is chosen by splitting the tree at different levels without any iteration.

2.1 K-Means Clustering

According to works of literature, the K-Means clustering algorithm has been identified as the best partitional clustering algorithm, and it is highly scalable and relatively fast. K-Means starts by selecting K representative points as initial centroids. Each data point is assigned to the nearest centroid iteratively, using a proximity measure such as Euclidean distance, Manhattan distance, and Cosine Similarity. The centroid for each class is recalculated and reassigned at every iteration until the centroid no longer changes, thus reducing the Sum of Squared Errors (SSE) for a given set of centroids.

Given a dataset $D = \{x_1, x_2, x_3, x_4, \dots, x_N\}$ of N points. c_k is the centroid of the cluster c_k. C_k is a set of the centroids from all clusters

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - c_k\| \quad (1)$$

$$C_k = \frac{\sum_{x_i - c_k}(x_i)}{|c_k|} \quad (2)$$

The method requires that the number of clusters, k, be preselected. The pre-selection has been identified as a lag as there is no prior knowledge of the dataset. To mitigate this, an adaptive K-Means can be adopted. The method automatically sets k according to the input dataset using any of the clustering indicators to determine the optimal number of clusters.

2.2 Hierarchical Clustering

The disadvantages identified with partitional clustering methods led to the development of flexible algorithms such as Hierarchical algorithms. One of the major issues identified with partitional clustering methods is the need for the user to predefine specific metrics that are non-deterministic in nature such as the number of clusters. Hierarchical algorithms can be generally classified into (i) Agglomerative methods (ii) Divisive methods. The agglomerative starts by making every data point a cluster; at the bottom level. Two clusters are merged at a time to build a bottom-up hierarchy of the clusters. The merging continues until there is only one cluster.

On the other hand, the Divisive method starts with all the data points in a singleton cluster and continues to split into two groups; resulting in a top-down cluster hierarchy [8]. The major peculiarity of a Hierarchical algorithm is that it allows the cutting of the hierarchy at any given level, yielding the clusters correspondingly. Hence, the number of clusters does not need to be predefined. Agglomerative Hierarchical clustering was adopted for this work following the assertion of [9] that Divisive methods are not always used for clustering load data in a power system due to the complexity of the algorithm and the high computational time and sensitivity to noise [8]. Ward's criterion was used in this work. It uses K-means squared error criterion to determine the distance during clusters' evaluation for merging.

2.3 Data Analysis

The data used in the research is collected from the Smart Energy Research Laboratory (SERL) – a research laboratory funded by the Tertiary Education Trust Fund (TETFund), Nigeria. The data contains energy data collected using energy monitors developed by SERL to harvest energy parameters from consumers for research purposes. The energy monitor collects two (2) energy data points every minute from a building and sends the data obtained through the wireless medium to the SERL database in the cloud. This paper uses three (3) month data collected from a residential building in Osogbo, Osun state – 'OSU' and a bakery factory situated Federal University of Technology, Akure, Ondo state - 'OND' were analyzed. 11870 and 10008 data points calibrated periodically in kWh were acquired from 'OND' and 'OSU', respectively.

2.3.1 Data cleansing and filtering

Raw energy data has the propensity to include missing points or outliers; this tends to corrupt the dataset and

influences the clustering result. Two (2) major steps are involved in cleaning a dataset from such aberrance – (i) Identification (ii) Treatment. No missing data point was identified from the two datasets (i.e., 'OSU' and 'OND'), hence, no treatment was required. However, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was adopted to determine the outliers in the datasets. DBSCAN is an unsupervised learning algorithm. The algorithm allows the tuning of two parameters: (i) the minimum distance between two data points for them to be a cluster in the same neighborhood – 'eps' and (ii) the minimum number of samples in a neighborhood for a data point to qualify as a core point – 'min_sample'. On tuning the 'eps' and 'min_sample' to 0.5 and 2 respectively, no outliers were found in 'OSU' but 3 were found in 'OND'. The three (3) outliers were removed from the total set as the calculated differences between each outlier, and its neighbors are unrealistic. In addition, a discard was employed because the number of identified outliers was few.

2.3.2 Data normalization

Energy data collected are complex and diverse. [10] noted that this can affect the clustering algorithm, and the original energy data may not satisfy Gaussian distribution [11]. Hence, the need for data normalization. Data Normalization sets the data points within a range. The z-scores normalization is identified to be poor by works of literature [12]. Hence, the unity-based normalization expressed in (3) was adopted for this work.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3)$$

Figure 1 shows the dispersion between 'OSU' and 'OND' data sequence before and after normalization. It is evident that the profiles of the two datasets are quite similar after normalization

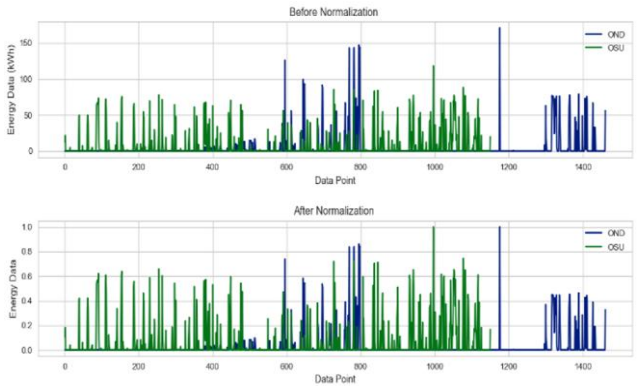

Figure 1: Normalization of the hourly load curve for' OND' and 'OSU'

2.3.3 Selection of clustering technique and number of clusters

To select the optimal number of k for 'OND', the elbow method shown in Figure 2 was considered. The Elbow method requires drawing a line plot between SSE (Sum of Squared Errors) and the number of clusters. The SSE value indicates the overall SSE obtained from each cluster. In this approach, the optimal number of clusters is considered the value of k at the "elbow" i.e., the point after which the distortion/inertia starts decreasing linearly. Following this rule, four (4) can be preselected as the optimal number of clusters for 'OND' and 'OSU'.

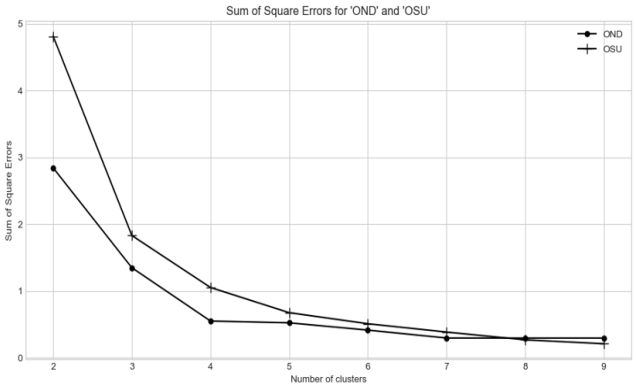

Figure 2: Sum of Square Errors plot of 'OND' and 'OSU'

2.3.4 Clustering indicators

The DB index and Silhouette Index evaluate the quality of clustering using the information embedded in the data. A higher value of SI means the clusters are well separated, while a lower DB index indicates a better clustering result [8]. These parameters are used in this work in the selection of the best clustering technique. Shown in Table 1 are the mathematical definitions of the clustering indicators.

Table 1: Clustering Indices Definitions

Clustering Indices	Definitions		Best Clustering Result
DBI	$\frac{1}{K}\sum_{i=1}^{K}\left(\frac{d\prime(L^{(i)}+d\prime(L^{(i)}))}{d(r^{(i)},r^{(j)})}\right)\ i \neq j$	(4)	Minimum
SI	$\frac{b(i)-a(i)}{max\{a(i),b(i)\}}$	(5)	Maximum

K is the total number of clusters, $L^{(i)}$ is the set of objects in cluster i, $r^{(i)}$ is the centroid of cluster i, d is the sum of the distance between objects in the cluster and the cluster centroid, $d'(L^{(i)})$ is the geometric mean of the inter- distance between objects in $L^{(i)}$, $d(r^{(i)}, r^{(j)})$ is the distance between centroids of cluster i and j, $a\ (i)$ is the mean of intra cluster distance, $b\ (i)$ is the mean nearest-cluster distance for each sample.

2.3.5 'OND' clustering technique

From the results obtained from Table 2 and Table 3 and the principle of selecting an optimal number of clusters, the Minimum value of DBI and Silhouette score infer the best number of clusters to choose. Selecting two (2) as the optimal number of clusters would have been the best decision, but such could not be settled due to the steepness observed in the 'OND' DBI plot. From the plot of SSE against the number of clusters shown in Figure 2, the "elbow" joint is found at the point 4. Hence, in selecting the best clustering technique for 'OND', the minimum value of the DBI plot and the maximum value of the SI plot for cluster four (4) were used to ascertain the best technique. Hierarchical clustering gave the minimum value (0.351152) from the DBI indices and the maximum value from the SI indices (0.944325). Hence, the adoption of Hierarchical clustering for 'OND'.

Table 2: 'OND' DBI Clustering Indices with varying cluster number

	DBI	
Number of Clusters	**Hierarchical Clustering**	**K-means Clustering**
2	0.445625	0.315271
3	0.355829	0.278410
4	0.351152	0.349489
5	0.410934	0.295307
6	0.390327	0.346389
7	0.369423	0.299564
8	0.383248	0.389464
9	0.352145	0.357671

Table 3: 'OND' SI Clustering Indices with varying cluster number

	SI	
Number of Clusters	**Hierarchical Clustering**	**K-means Clustering**
2	0.941991	0.952603
3	0.946207	0.951532
4	0.944325	0.939319
5	0.935536	0.942375
6	0.937957	0.936819
7	0.939822	0.936055
8	0.937520	0.942757
9	0.937100	0.942337

2.3.6 'OSU' clustering technique

From the clustering indices shown in Table 4 and Table 5, 'OSU' showed a gentle decline and increase in the values of SI and DBI, respectively, as the k increases from 2 to 9. Furthermore, having preselected four (4) as the optimal number of clusters, the clustering indices obtained from DBI and SI plots were used. Hierarchical clustering gave a minimum value of 0.406111 from the DBI indices, while K-Means clustering gave the maximum value of 0.872986. To adjudge the best clustering technique, the differences between the DBI and SI values at cluster number 4 were observed. DBI indices showed a wider difference (0.042012) than that of SI (0.013103). Hence, DBI was used to adjudge. From the results, hierarchical clustering will be the best clustering technique to be adopted for 'OSU'.

Table 4: 'OSU' DBI Clustering Indices with varying cluster number

	DBI	
Number of Clusters	**Hierarchical Clustering**	**K-means Clustering**
2	0.438949	0.330031
3	0.413650	0.408763
4	0.406111	0.448123
5	0.418220	0.507681
6	0.487274	0.469537
7	0.419405	0.476662
8	0.433933	0.474027
9	0.418471	0.388977

Table 5: 'OSU' SI Clustering Indices with varying cluster number

	SI	
Number of Clusters	**Hierarchical Clustering**	**K-means Clustering**
2	0.865629	0.878978
3	0.877208	0.876575
4	0.859883	0.872986
5	0.872043	0.864095
6	0.872146	0.880194
7	0.872656	0.880300
8	0.877179	0.882740
9	0.878306	0.882395

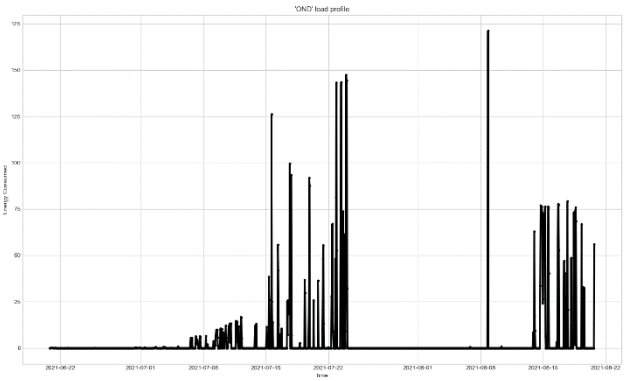

Figure 3: Load Profile of 'OND'

3.0 RESULTS AND DISCUSSIONS

From Figure 4, 'OSU' has stochastic consumption behavior being a residential building. A maximum peak consumption of 117kW was obtained for the 3 months while 'OND' has a peak demand of about 175kW, being a commercial building. However, from the analysis, 'OSU' has steady power supply than 'OND', a setback that accounts for the flat line in 'OND' profile plot. Figure 3 and Figure 4 only summarized what is obtainable in 'OND' and 'OSU'. However, the cluster results obtained from the

hierarchical clustering done will be employed. Both 'OND' and 'OSU' have been clustered into 4 (four) different clusters.

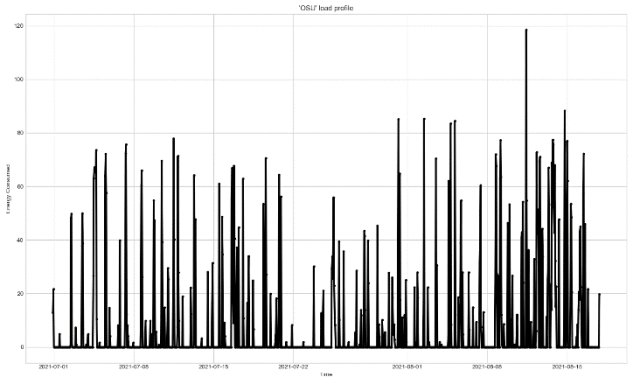

Figure 4: Load Profile of 'OSU'

3.1 'OND' Clustering Results

Using the Hierarchical clustering technique, the dendrogram generated from 'OND' were split at point 1.1, in the bid to have 4 clusters. Based on Figure 5, the pattern of cluster 1 is very irregular, and the range of the energy consumed is between 0kWh and 32kWh. The cluster also captures periods when the building is not supplied with power. The average consumption of 31.59kWh was obtained from the analysis. Hence, the consumption infers a minimal consumption which can be interpreted as periods when usage is within the sphere of essential appliances. The load demand during these periods is spread across the 24-hour of each day. Cluster 2 shows less stochastic consumption that spans across the period of collection.

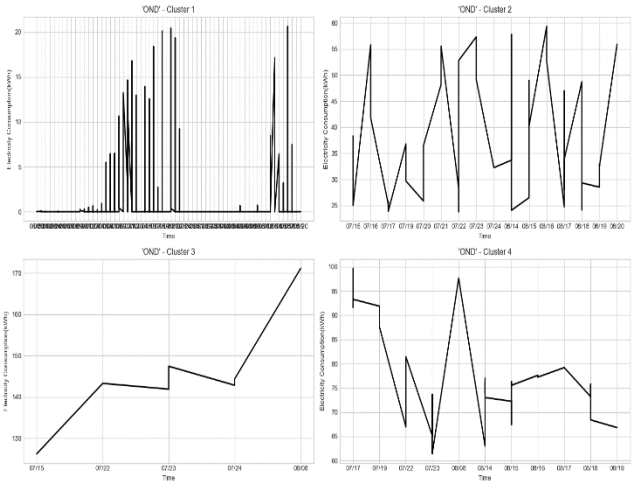

Figure 5: 'OND' clusters' result

On further analysis, with load demand as high as 60kWh, the pattern is consistent between 08:00 and 22:00 each day. This pattern correlates strongly with what was obtained in cluster 4, maintaining an average consumption of 142kWh. As shown in Figure 6, the demand continues to increases up to 22:00. This can be interpreted as the period of work at the bakery, when heavy pieces of equipment are used. Cluster 3 showed a sharp increase in energy demand from 100kWh to 175kWh between 15th July and 8th August. This could be as a result of activities or occurrences on campus capable of raising demands on the product turnout of the factory. Generally, the bakery makes a high demand, but much more between the 8:00 and 22:00 each day as inferred by cluster 2 and 4.

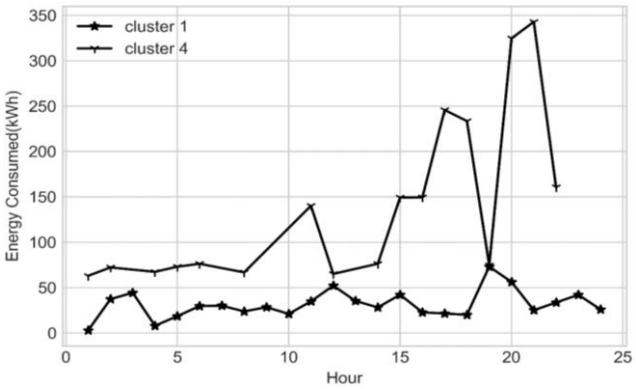

Figure 6: 'OND' cluster 1 and 4 cumulative result hourly

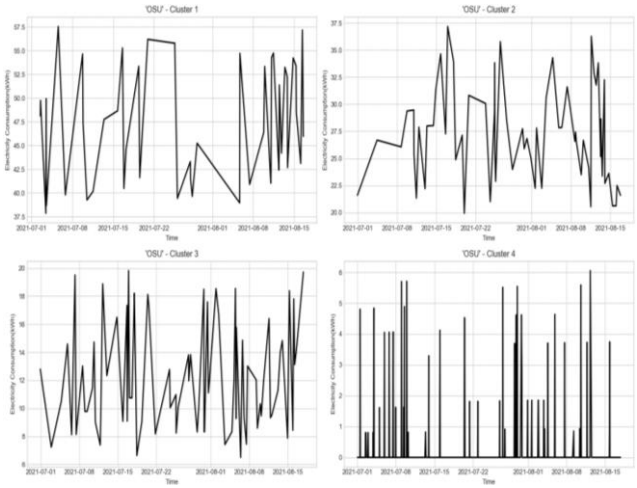

Figure 7: 'OSU' clusters result

3.2 'OSU' Clustering Results

Using the Hierarchical clustering technique, the dendrogram generated from 'OND' was split at 1.0 in the bid to have 4 clusters. Based on Figure 7, 'OSU' had a more regular power supply than 'OND' due to the geographical location difference. Analyzing the overall energy consumption, high consumption values were recorded on weekends, a significantly low consumptions are recorded on 'Tuesdays' and 'Thursdays'. Being a residential building, this can be

interpreted as the absence of humans in the building. Furthermore, in similitude of the result obtained from 'OND's cluster 1, much variableness was observed in cluster 1 of 'OSU'. However, the maximum energy consumption is lesser than 'OND' (average value of 17.9kWh). This infers the absence of heavy-duty equipment.

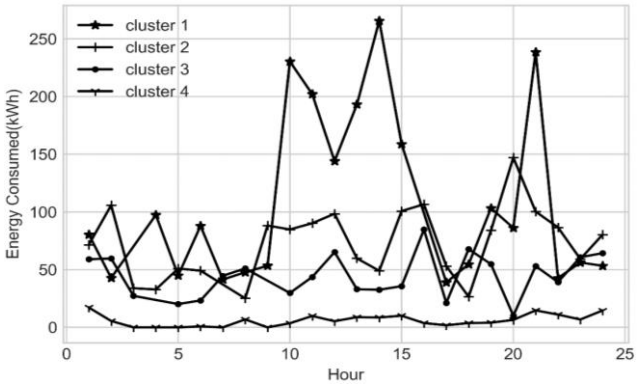

Figure 8: 'OSU' clusters' cumulative results hourly

The variableness of cluster 2, 3, 4 could not be interpreted due to the limited volume of the data. However, based on Figure 8, a typical pattern was found in all the clusters. The demand rises from 9:00 up to 17:00 through all the clusters. Although the concerned building is residential, this observation saliently reveals activities in the building at such period, which is unusual of a residential building. The result thus indicates the presence of commercial activities in the building.

4.0 CONCLUSIONS AND FUTURE WORKS

Electrical load profiles can be clustered using different kinds of clustering techniques. According to works of literature, K-Means has been identified as the most effective. However, such a conclusion was not obtainable from this work using Davies-Bouldin Index and Silhouette Index as clustering indicators. Hierarchical clustering was found to be best for the two-dataset analysis. Therefore, it is evident that the choice of clustering technique is highly dependent on the intrinsic nature of the dataset. From clusters obtained from 'OND' and 'OSU', differentiating between a commercial building and a residential building is easy using data mining techniques. The cluster 1 results obtained from 'OSU' and 'OND' show minimal but highly irregular consumption periods. In contrast, cluster 2 and cluster 4 of 'OND' reveal the active period of the commercial building(bakery). The results from clusters 1,2,3,4 all show an unusual consumption pattern in a building that appears to be a residential building.

Data mining brings out hidden information from data; electrical load profiling is one of its applications. Its strength has been proved in this work. This technique should be employed for other applications such as energy theft detection and other energy management programs that benefit utilities and power providers.

5.0 ACKNOWLEDGMENT

The authors would like to thank the Smart Energy Research Laboratory (SERL) for providing the data used for analysis. This research was funded by TETFund Research Fund Grant 2019 (TETFund/DR&D-CE/NRF/2019).

REFERENCES

[1] Kane, S, N., Mishra, A., and Dutta, A. K. "Electrical Load Profile Analysis Using Clustering Techniques", in *Journal of Physics: Conference Series*, 2016, vol. 755, no. 1, doi: 10.1088/1742-6596/755/1/011001.

[2] Zhou, K., Fu, C., and Yang, S. "Big data driven smart energy management: From big data to big insights", vol. 56, no. 2016, pp. 215–225, 2020, doi: 10.1016/j.rser.2015.11.050.

[3] Liu, X., Ding, Y., Tang, H., and Xiao, F. "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data", *Energy Build.*, vol. 231, no. xxxx, p. 110601, 2021, doi: 10.1016/j.enbuild.2020.110601.

[4] Oseni, M. O. "An analysis of the power sector performance in Nigeria", *Renew. Sustain. Energy Rev.*, vol. 15, no. 9, pp. 4765–4774, 2011, doi: 10.1016/j.rser.2011.07.075.

[5] Wang, Y., Chen, Q., Kang, C., Zhang, M., Wang, K., and Zhao, Y. "Load Profiling and Its Application to Demand Response: A Review", *Tsinghua Sci. Technol.*, vol. 20, no. 2, pp. 117–129, 2015, doi: 10.1109/TST.2015.7085625.

[6] Gunsay, M., Bilir, C., and Poyrazoglu, G. "Load Profile Segmentation for Electricity Market Settlement", in *2020 17th International Conference on the European Energy Market (EEM)*, 2020, pp. 1–5, doi: 10.1109/EEM49802.2020.9221889.

[7] Il Kim, Y., Ko, J. M., and Choi, S, H. "Methods for generating TLPs (Typical Load Profiles) for smart grid-based energy programs", *IEEE SSCI 2011 - Symp. Ser. Comput. Intell. - CIASG 2011*

2011 IEEE Symp. Comput. Intell. Appl. Smart Grid, pp. 49–54, 2011, doi: 10.1109/CIASG.2011.5953331.

[8] C. C. A.; Chandan K. Reddy, *DATA Algorithms and Applications*. Minnesota, U.S.A., 2014.

[9] Khan, Z. A., Jayaweera, D., and Alvarez-alvarado, M. S., "A novel approach for load profiling in smart power grids using smart meter data", *Electr. Power Syst. Res.*, vol. 165, no. August, pp. 191–198, 2018, doi: 10.1016/j.epsr.2018.09.013.

[10] Viegas, J. L., Vieira, S. M., Melício, R., Mendes, V. M. F., and Sousa, J. M. C. "Classification of new electricity customers based on surveys and smart metering data", *Energy*, vol. 107, no. 2016, pp. 804–817, 2016, doi: 10.1016/j.energy.2016.04.065.

[11] Lu, S., Lin, G., Liu, H., Ye, C., Que, H., and Ding, Y. I. "A Weekly Load Data Mining Approach Based on Hidden Markov Model", *IEEE Access*, vol. 7, pp. 34609–34619, 2019, doi: 10.1109/ACCESS.2019.2901197.

[12] Zhan, S., Liu, Z., Chong, A., and Yan, D. "Building categorization revisited: A clustering-based approach to using smart meter data for building energy benchmarking", *Appl. Energy*, vol. 269, no. February, 2020, doi: 10.1016/j.apenergy.2020.114920.