Load Profiling of a residential Building and a commercial building using Clustering Techniques

Olawumi Abayomi Ebenezer

Abstract

Data mining is an auspicious tool used in processing energy data collected from energy consumers. The knowledge derived from energy data is very germane in the formulation of various demand side management programs. This paper uses clustering techniques in segmenting the various consumption behaviors of a residential building and a commercial building in different geographical locations. The two (2) 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. In this work, the optimum number of clusters was selected using the Davies-Bouldin Index (DB) Index and Silhouette Index (SI). Also, Hierarchical clustering was found to be the most appropriate clustering.

Keywords – Load Profiling, Clustering, data mining, machine learning

1.0 Introduction

Analysis of consumption behavior of energy consumers is very paramount to the planning and development of smart grid. In smart grid, smart meters are used to collect and store energy parameters at different resolution. The data collected from different geographical locations are transmitted in real-time to a central location where intelligent decisions are taken and reverted back to the smart meters(Kane et al., 2016). In a modern power grid system, the energy data and other related information are instrumental for applications such 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 pattern is very important in the formulation of tariffs systems and also 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(Zhou et al., 2020), a more intelligent approach is needed in the analysis of 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. The approach is known to be more effective than statistical methods as it prevents loss of energy estimates(Kane et al., 2016)(Liu et al., 2021). In this paper, clustering techniques will be used to analyze the consumption patterns of energy consumers. In Nigeria, residential and commercial sectors represents about 80% of the total electricity demand (Oseni, 2011). However, among the various clustering techniques, K-Means and Hierarchical Clustering are well known for high performance(Wang et al., 2015)(Gunsay et al., 2020). 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(Y. Il Kim et al., 2011). 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 Clustering Techniques

Clustering is an unsupervised learning algorithm that intelligently classify objects with similar attributes in a group. The algorithm is capable of bringing out hidden similarities from a sparse data. It is widely employed in various disciplines such as image processing, pattern recognition, etc. In this work, the technique is used to group the various 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(Kim et al., 2011). The partitonal 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.

K-Means clustering

According to works of literature, K-Means clustering algorithm have been identified as the best partitional clustering algorithm, 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.

$c_k =$ Centroid of cluster C_k.

C_k = $\{C_1, C_2, C_3, C_4, \dots, C_k\}$.

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left\| x_i - c_k \right\|^2 \qquad (1)$$

$$C_k = \frac{\sum_{x_i \in C_k} (x_i)}{|c_k|} \qquad (2)$$

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

Agglomerative 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 certain metrics that are non-deterministic in nature e.g. 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 point in a singleton cluster and continue to split into two group; resulting to a top-down hierarchy of cluster(Reddy, 2014). 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 Khan et al., (2018) 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(Reddy, 2014). Ward’s criterion was used in this work. It uses K-means squared error criterion to determine the distance during clusters’ evaluation for merging.

3. 0 Methodology

A. Data Collection

The data used in the research is collected from the Smart Energy Research Laboratory (SERL) – a research group funded by the Tertiary Education Trust Fund (TETFund), Nigeria. The data contains energy data collected using energy monitors designed by SERL. The energy monitor averagely collects two (2) energy data points every minute from a building and sends the data acquired through the cloud to SERL database. In this paper, a 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.

B. Data Cleansing and Filtering

Raw energy data has the propensity of containing missing points or outliers, this has the tendency of corrupting the dataset and influence the clustering result. Two (2) major steps are involved in order clean 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, in order to identify the outliers in the datasets, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was adopted. 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 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 it neighbors are unrealistic. In addition, a discard was employed based on the fact that the number of identified outliers were few.

C. Data Normalization

Energy data collected are complex and diverse in nature. Viegas et al., (2016) noted that this can affect the clustering algorithm and the original energy data may not satisfy Gaussian distribution(Lu et al., 2019). Hence, the need for data normalization. Data Normalization sets the data points within a range. The z-scores normalization has been identified to be poor by works of literature(Zhan et al., 2020). Hence, the unity-based normalization expressed in equation 1 was adopted for this work.

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

Where x represents the original energy data sequence, min(x) represent the minimum energy data sequence, the max(x) represents the maximum energy data sequence and x' represents the normalized energy data sequence.

Figure 1.0 shows the dispersion between the data sequence of ‘OSU’ and ‘OND’ before and after normalization. It is obvious that the profiles of the two dataset are quite similar after normalization.

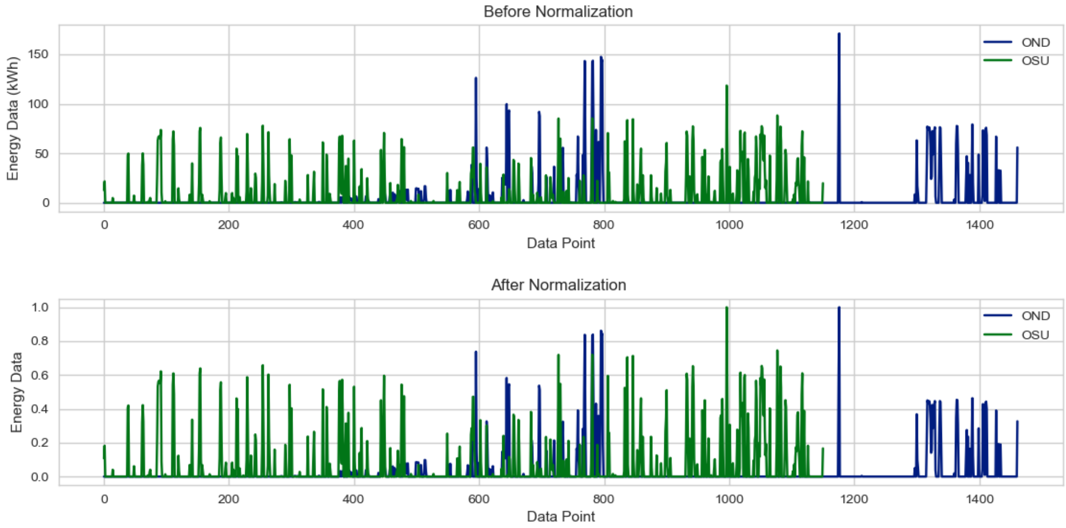


Figure 1. Normalization of the hourly load curve for ’OND’ and ‘OSU’

D. SELECTION OF CLUESTERING TECHNIQUE AND NUMBER OF CLUSTERS

In order to select the optimal number of k for ‘OND’, the elbow method shown in figure 4 was considered. 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 as the value of k at the “elbow” i.e. the point after which the distortion/inertia start decreasing in a linear fashion. As shown in figure 3.1 and following this rule, 4 can be preselected as the optimal number of clusters for ‘OND’ and ‘OSU’

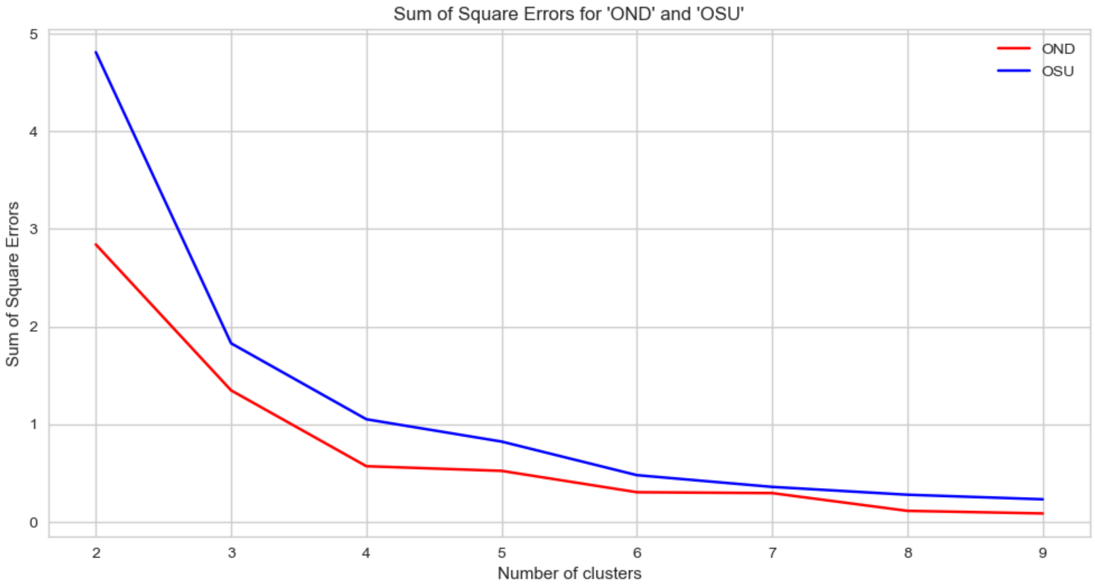


Figure 3.1: Sum of Square Errors plot of ‘OND’ and ‘OSU’

CLUSTERING INDICATORS

The DB index and Silhouette Index evaluate the quality of clustering using the information embedded in the data. Higher value of SI means the clusters are well separated while lower DB index indicates a better clustering result (Reddy, 2014). These parameters are used in this work in the selection of the best clustering technique. Shown in table 1 is the mathematical definitions of the clustering indicators.

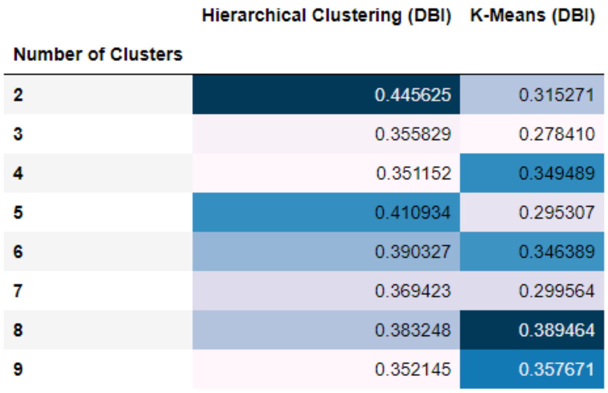
Table 1. Clustering Indices definitions

Clustering Index	*Definitions*		*Best Clustering Result (k)*
DBI	$\frac{1}{K}\sum_{i=1}^{k}\left(\frac{d'\left(L^{(i)}+d'\left(L^{(i)}\right)\right)}{d\left(r^{(i)},r^{(j)}\right)}\right)\ i\ \neq j$	(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, $d'\left(L^{(i)}\right)$ is the geometric mean of the inter-distance between objects in $L^{(i)}$, $d\left(r^{(i)},r^{(j)}\right)$ 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.

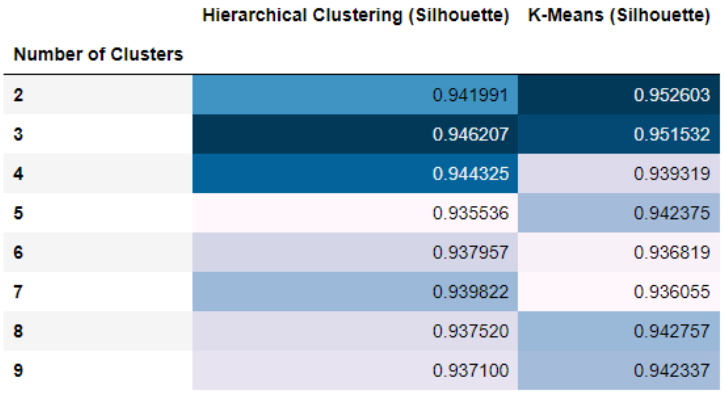
‘OND’ CLUSTERING TECHNIQUE

Table 3.1 ‘OND’ DBI Clustering Indices with varying cluster number



Number of Clusters	Hierarchical Clustering (DBI)	K-Means (DBI)
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 3.2 ‘OND’ SI Clustering Indices with varying cluster number

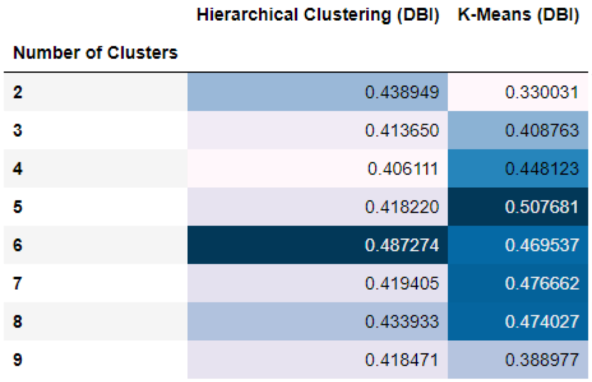


Number of Clusters	Hierarchical Clustering (Silhouette)	K-Means (Silhouette)
2	0.941991	0.952603
3	0.946207	0.951532
4	0.944325	0.939319
5	0.935536	0.942375
6	0.937957	0.936819
7	0.939822	0.936055
8	0.937520	0.942757
9	0.937100	0.942337

From the results obtained from table 3.1 and table 3.2 and the principle of selecting optimal number of cluster, Minimum value of DBI and Silhouette score infer the best number of cluster to select. Selecting 2 as the optimal number of cluster would have been the best decision but such could not be settled for due to the steepness observed in ‘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 to adopt for ‘OND’, the minimum value of the DBI plot and the maximum value of the SI plot for cluster 4 was used in ascertaining the best technique. *Hierarchical clustering* gave the minimum value (0.351152) from the DBI indices and also gave the maximum value from the SI indices (0.944325). Hence, the adoption of Hierarchical clustering for ‘OND’.

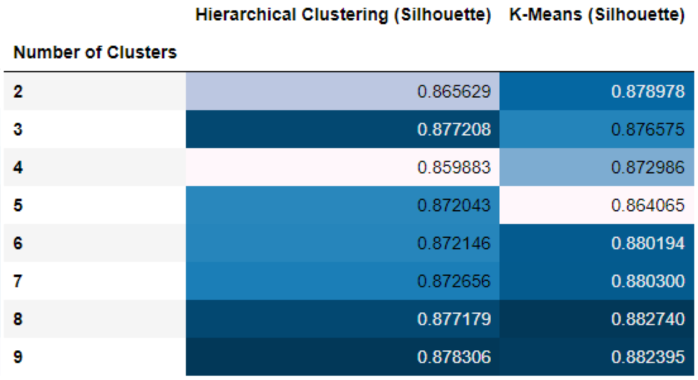
‘OSU’ CLUSTERING TECHNIQUE

Table 3.3 ‘OSU’ DBI Clustering Indices with varying cluster number



Number of Clusters	Hierarchical Clustering (DBI)	K-Means (DBI)
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 3.4 ‘OSU’ SI Clustering Indices with varying cluster number



Number of Clusters	Hierarchical Clustering (Silhouette)	K-Means (Silhouette)
2	0.865629	0.878978
3	0.877208	0.876575
4	0.859883	0.872986
5	0.872043	0.864065
6	0.872146	0.880194
7	0.872656	0.880300
8	0.877179	0.882740
9	0.878306	0.882395

From the clustering indices shown in table 3.3 and table 3.4. ‘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 4 as the optimal number of clusters, the clustering indices obtained from DBI plot and SI plot 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. In order 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’.

4.0 Findings and Discussions

From figure 4.1b, ‘OSU’ has stochastic consumption behavior being a residential building, It has a maximum peak consumption of about 117kW 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 which accounts for the flat line in ‘OND’ profile plot.

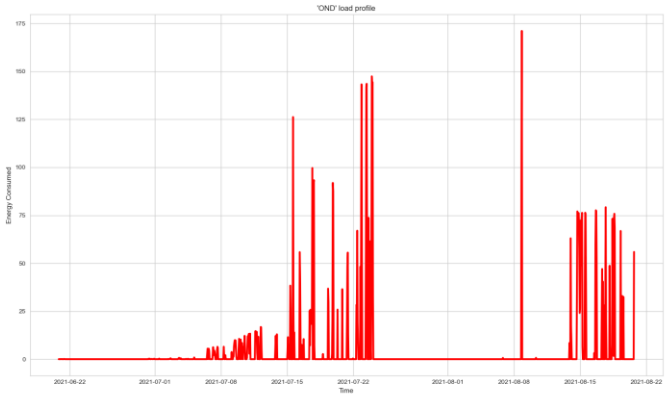


Figure 4.1a: Load Profile of ‘OND’

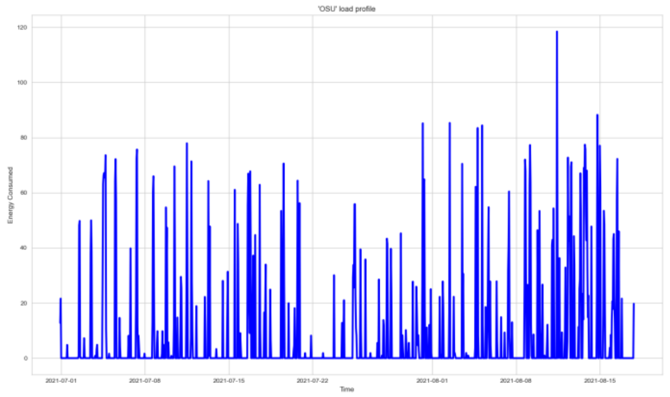


Figure 4.1b: Load Profile of ‘OSU’

Figure 4 only gave a summarized information of what is obtainable in ‘OND’ and ‘OSU’. However, in order to further investigate the consumption behaviors, the clusters results obtained from the hierarchical clustering done will be employed. Both ‘OND’ and ‘OSU’ have been clustered into 4 (four) different clusters.

‘OND’ CLUSTERING RESULTS

Using Hierarchical clustering technique and selecting the number of cluster as 4, the result of clustering shown in the dendrogram illustrated in figure 4.1. The dendrogram generated from ‘OND’ were split at 1.1 in the bid to have 4 clusters.

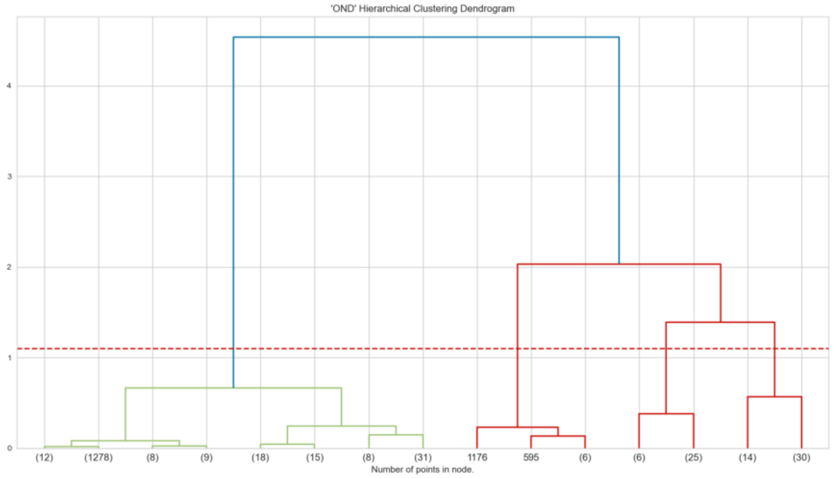


Figure 4.2. Dendrogram of ‘OND’

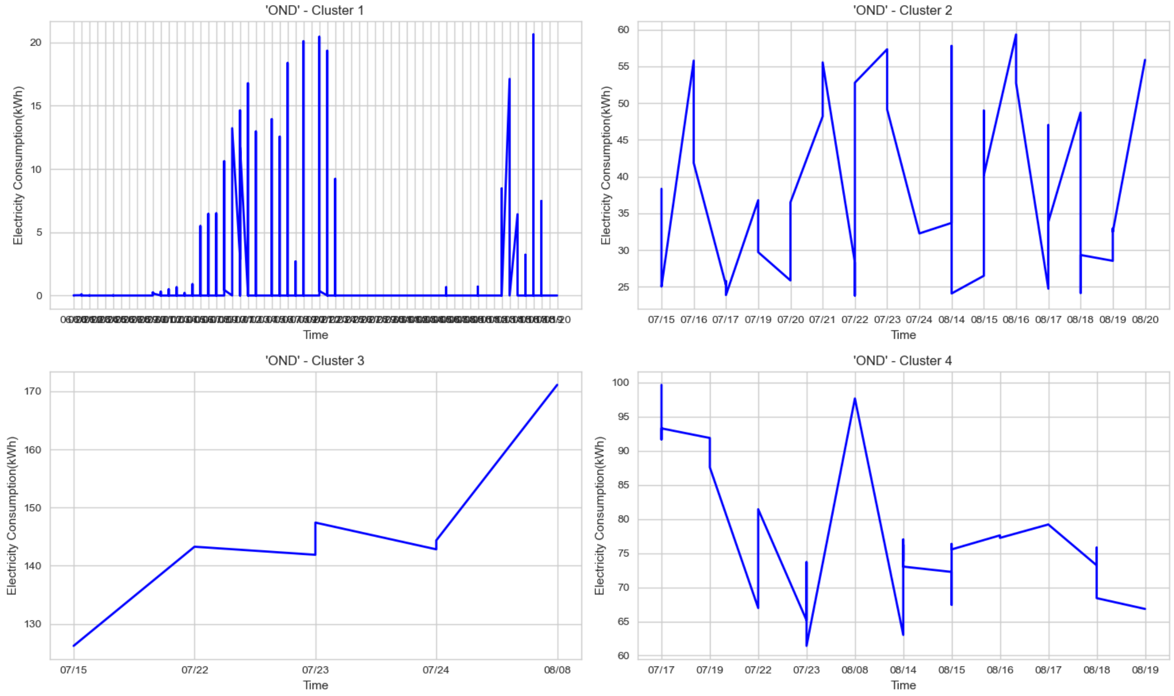


Figure 4.3. ‘OND’ clusters’ result

Based on figure 4.3, 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 analysis. Hence, the consumption infers a minimal consumption which can be interpreted as periods when usage is within the sphere of basic 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. 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, both maintaining an average consumption of 142kWh. As shown in figure 4.4, 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, a high demand is made by the bakery, but much more between the 8:00 and 22:00 each day as inferred by cluster 2 and 4.

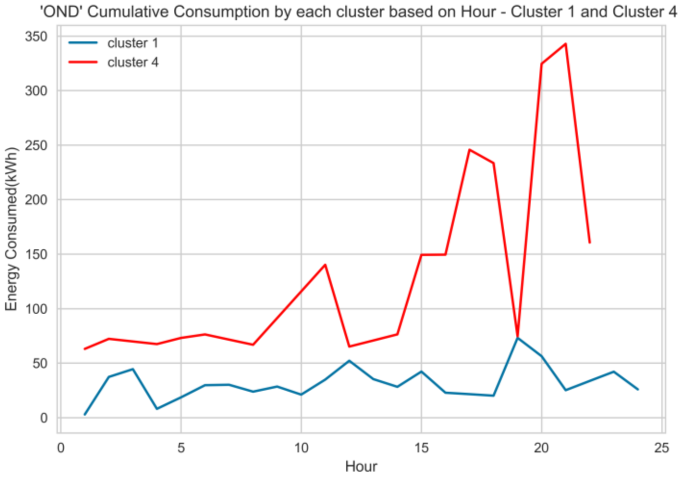


Figure 4.4. ‘OND’ cluster 1 and 4 cumulative result based on hour

‘OSU’ CLUSTERING RESULTS

Using Hierarchical clustering technique and selecting the number of cluster as 4, the result of clustering shown in the dendrogram illustrated in figure 4.1. The dendrograms generated from ‘OND’ were split at 1.0 in the bid to have 4 clusters.

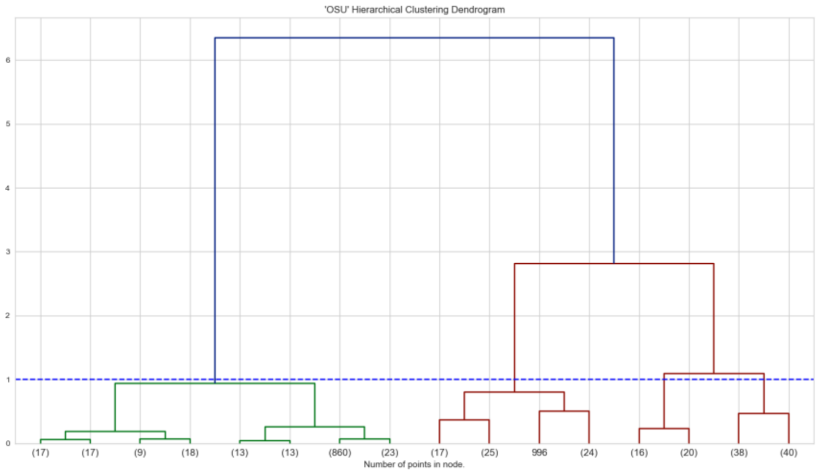


Figure 4.4. Dendrogram of ‘OSU’

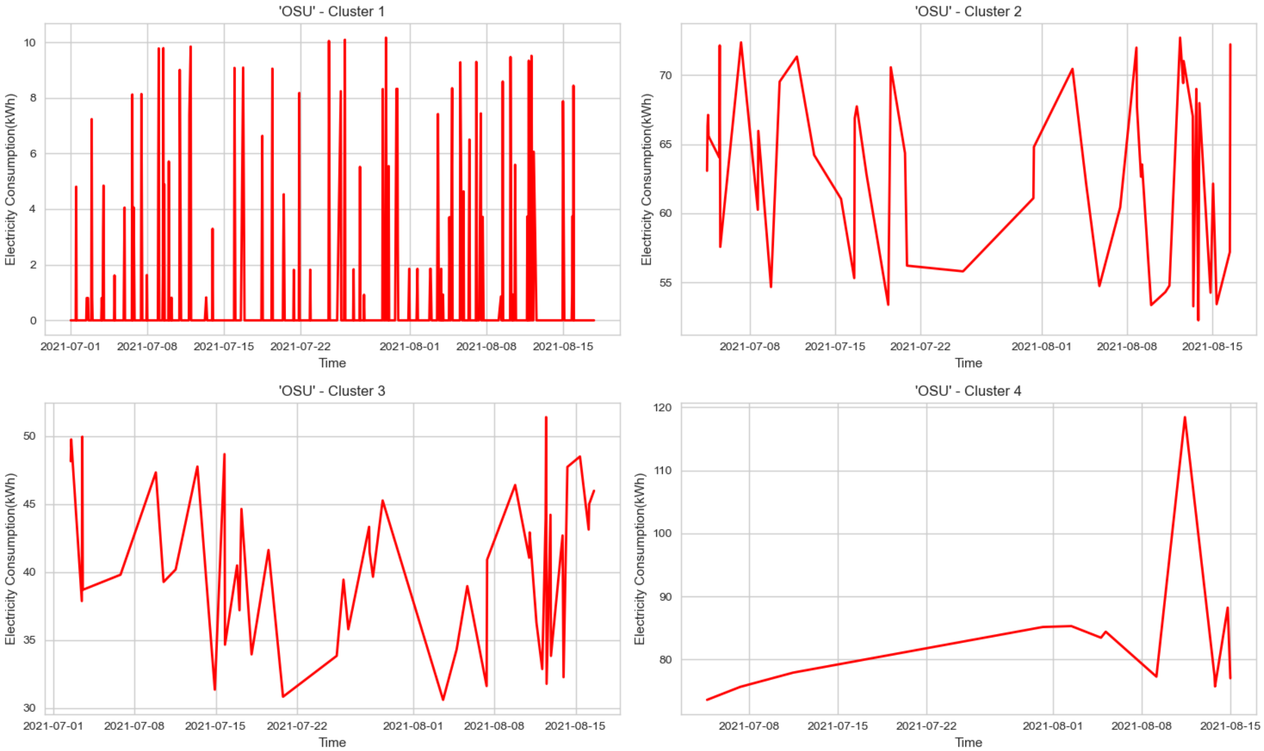


Figure 4.5. ‘OSU’ cluster result

Based on figure 4, ‘OSU’ had more regular power supply than ‘OND’; owing to the difference in geographical location. 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. So much variableness was found in cluster 1 of ‘OSU’. However, the maximum energy consumption is lesser than that of ‘OND’ (average value of 17.9kWh), this infers the absence of heavy duty equipment. The variableness of cluster 2, 3, 4 could not be interpreted due to the limited volume of the data. However, based on figure 4.6, a common pattern was found in all the clusters. The energy demand rise from 9:00 up to 17:00 through all the clusters. Although, the concerned building is residential, this observation saliently reveals the presence of activities in the building at such period, which is unusual of a residential building. The result thus reveals the presence of commercial activities in the building.

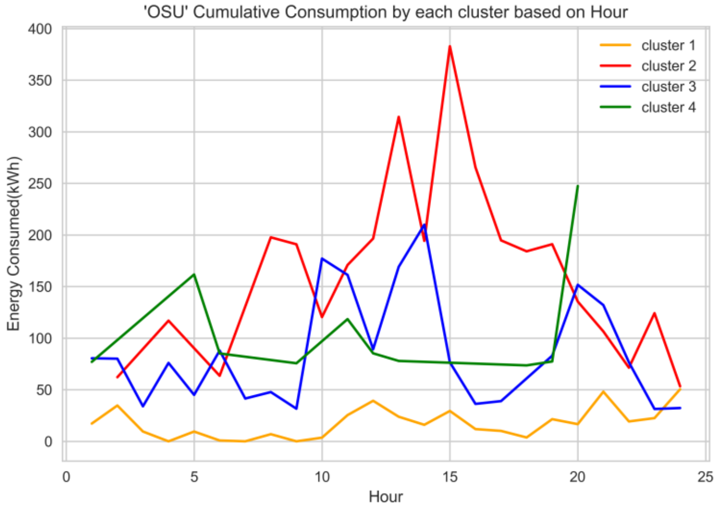


Figure 4.6. ‘OSU’ clusters’ cumulative results based on hour

5. 0 CONCLUSIONS

Electrical load profiles can be clustered using different kinds of clustering techniques. According to works of literature, K-Means has been identified as most effective, however, such 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. It is therefore obvious 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 made easy using data mining techniques. The cluster 1 results obtained from ‘OSU’ and ‘OND’ show the periods of minimal but highly irregular consumption, while cluster 2 and cluster 4 of ‘OND’ reveals the active period of the commercial building(bakery). The results from clusters 1,2,3,4 all reveals an unusual consumption pattern in a building that appears to be a residential building.

6. 0 RECOMMENDATIONS

Data mining is capable of bringing out hidden information from data, electrical load profiling is one of its application. Its strength has been proved in this work. This technique should be employed for other kinds of application such as energy theft detection and other energy management programs that can be more beneficial to utilities and power providers.

References

Gunsay, M., Bilir, C., & Poyrazoglu, G. (2020). Load Profile Segmentation for Electricity Market Settlement. *2020 17th International Conference on the European Energy Market (EEM)*, 1–5. https://doi.org/10.1109/EEM49802.2020.9221889

Kane, S. N., Mishra, A., & Dutta, A. K. (2016). Electrical Load Profile Analysis Using Clustering Techniques. *Journal of Physics: Conference Series*, *755*(1). https://doi.org/10.1088/1742-6596/755/1/011001

Khan, Z. A., Jayaweera, D., & Alvarez-alvarado, M. S. (2018). A novel approach for load pro filing in smart power grids using smart meter data. *Electric Power Systems Research*, *165*(August), 191–198. https://doi.org/10.1016/j.epsr.2018.09.013

Kim, Y. Il, Ko, J. M., & Choi, S. H. (2011). Methods for generating TLPs (Typical Load Profiles) for smart grid-based energy programs. *IEEE SSCI 2011 - Symposium Series on Computational Intelligence - CIASG 2011: 2011 IEEE Symposium on Computational Intelligence Applications in Smart Grid*, 49–54. https://doi.org/10.1109/CIASG.2011.5953331

Kim, Y.-I., Kang, S.-J., Ko, J.-M., & Choi, S.-H. (2011). A Study for Clustering Method to generate Typical Load Profile for Smart Grid. *8th International Conference on Power Electronics - ECCE Asia*, 1102–1109.

Liu, X., Ding, Y., Tang, H., & Xiao, F. (2021). A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data. *Energy and Buildings*, *231*(xxxx), 110601. https://doi.org/10.1016/j.enbuild.2020.110601

Lu, S., Lin, G., Liu, H., Ye, C., Que, H., & Ding, Y. I. (2019). A Weekly Load Data Mining Approach Based on Hidden Markov Model. *IEEE Access*, *7*, 34609–34619. https://doi.org/10.1109/ACCESS.2019.2901197

Oseni, M. O. (2011). An analysis of the power sector performance in Nigeria. *Renewable and Sustainable Energy Reviews*, *15*(9), 4765–4774. https://doi.org/10.1016/j.rser.2011.07.075

Reddy, C. C. A. ;Chandan K. (2014). *DATA Algorithms and Applications* (2014 CRC Press Taylor & Francis Group (ed.)).

Viegas, J. L., Vieira, S. M., Melício, R., Mendes, V. M. F., & Sousa, J. M. C. (2016). Classification of new electricity customers based on surveys and smart metering data. *Energy*, *107*(2016), 804–817. https://doi.org/10.1016/j.energy.2016.04.065

Wang, Y., Chen, Q., Kang, C., Zhang, M., Wang, K., & Zhao, Y. (2015). Load Profiling and Its Application to Demand Response : A Review. *Tsinghua Science and Technology*, *20*(2), 117–129. https://doi.org/10.1109/TST.2015.7085625

Zhan, S., Liu, Z., Chong, A., & Yan, D. (2020). Building categorization revisited : A clustering-based approach to using smart meter data for building energy benchmarking. *Applied Energy*, *269*(February). https://doi.org/10.1016/j.apenergy.2020.114920

Zhou, K., Fu, C., & Yang, S. (2020). *Big data driven smart energy management : From big data to big insights*. *56*(2016), 215–225. https://doi.org/10.1016/j.rser.2015.11.050