Electrical Peak Load Clustering Analysis Using K-Means Algorithm and Silhouette Coefficient

Handrea Bernando Tambunan
Transmission and Distribution Department
PLN Research Institute
Jakarta, Indonesia
handrea.tambunan@pln.co.id

Dhany Harmeidy Barus
Transmission and Distribution Department
PLN Research Institute
Jakarta, Indonesia
barus@pln.co.id

Joko Hartono
Transmission and Distribution Department
PLN Research Institute
Jakarta, Indonesia
joko.hartono@pln.co.id

Aji Suryo Alam
Transmission and Distribution Department
PLN Research Institute
Jakarta, Indonesia
aji.alam@pln.co.id

Dimas Aji Nugraha
Transmission and Distribution Department
PLN Research Institute
Jakarta, Indonesia
dimas.aji@pln.co.id

Hakim Habibi Hidayatullah Usman
Transmission and Distribution Department
PLN Research Institute
Jakarta, Indonesia
hakim.habibi@pln.co.id

***Abstract*—Nowadays, data analysis widely used in many fields especially in engineering. Clustering is one of data analysis methods to organize the amount of data into groups with similarity characteristics. One powerful analysis method to learn information by grouping data is clustering algorithms. The clustering advantages for electrical power utilities is to learn load behavior and provide information for power plant operation and also generation cost. In this paper, a simulation concept is proposed for analysis of peak load data by K-means clustering algorithm based on historical dataset. The results show electrical peak loads clustering by K-means algorithm are optimum classified into three clusters. This cluster evaluated by silhouette scores which high, intermediate, and low load level interpretation. One cluster has centroid during January, June, and July are relatively lower than another cluster caused by Indonesia national holiday. This concept also evaluates the load level affected by Covid-19 pandemic condition.**

Keywords— clustering, Covid-19, k-means, pandemic, peak load, silhouette

I. INTRODUCTION

Clustering is an algorithm to organize the amount of data into meaningful groups with similarity or the same characteristic as much as possible. This algorithm is an analytical method in data analysis and also data mining to learn the information by grouping the data.

Data analysis with high dimensional data such as images, videos, text, measurement and more now widely used in many fields such as communication, biology, computer science, economic, epidemiology, psychology, business, and also engineering fields such as electrical power utilities [1]–[4]. In the present and future modern power system, the information from big data analytics has a great potential to be exploited for example from the demand side. It will provide the accurate information of customers consumption behaviour for system planning and operation by clustering approach. Accurate classification of electrical load is very essential for power plant operation and also generation cost.

Several clustering analysis for collection data points discovered in several papers [5]–[7]. There are many categories, algorithms, and evaluation indicators for clustering analysis with the dataset. One powerful clustering method is K-means as shown in some study [8]–[10]. These studies calculate the distance between data objects (centroid) and clusters by iteration process. The advantages of this method are relatively low time complexity and high computing efficiency. Several study literature determined the number of groups using silhouette coefficient to find the optimum number of clusters [11]–[13]. Different clustering techniques have been applied in some work for carrying out the electrical load pattern classification [14]–[16]. Study in [17] proposed an clustering analysis for electrical consumption pattern and load factor assessment in different days in particular zones using the K-means clustering algorithm.

In this paper, a proposed method to analyse the minimum peak load into clustered from historical data by K-means algorithm. This concept aims to classify the minimum peak load variation in multi years at different characteristic cluster especially in Covid-19 pandemic condition. This proposed concept help to identification the peak load with similar profile and characteristic especially in particular condition. The proposed approach is simulated on Indonesia-Jawa Bali power system.

II. DATA CLUSTERING

Clustering is an unsupervised learning that aims to group the dataset based on similarity and characteristic. Each cluster has a centroid as representation of the data pertaining to the cluster. Each data assigned to the nearest centroid. This paper simulates the historical data from Indonesia-Jawa Bali system during the 2014 to 2020 dataset.

A. Electrical Peak Load

Jawa Bali power system is one of the biggest electrical grids in Indonesia. The maximum, average, and also minimum monthly peak load from January 2014 to April 2020 specifically show in Fig. 1. Currently, the maximum peak load is approximately 27.9 GW in November 2019.

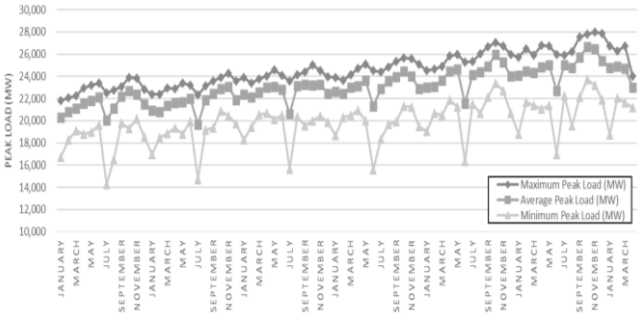

Fig. 1. Maximum, average, and minimum monthly electrical peak load.

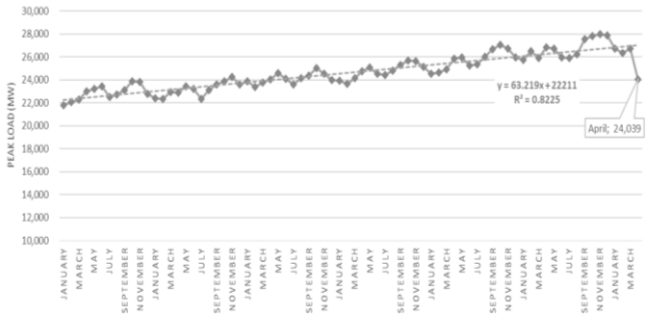

Fig. 2. Monthly maximum electrical peak load.

Jawa Bali system peak load constantly increases year by year in linear function as expressed in Fig. 2. The linear function is as follows f(x)=63.219x + 22,211. The lowest peak load in 2020 occurs in April. This condition is caused by Covid-19 pandemic so the electricity consumption tends to decrease [18].

B. K-means Algorithm

Basic construction of clustering with K-means algorithm used similarity and dissimilarity (distance) of the dataset. The illustration of K-means algorithm shows in Fig. 3 where (a) is input dataset, (b) is initial representatives (centroid) of the selected cluster and (c) compute the distance each data with smallest distance, repeat iteration (d) until final clustering convergence (e) [8].

Clustering algorithm categorized into 9 categories which are, hierarchy, fuzzy theory, distribution, density, graph theory, grid, fractal theory, model, and also partition [5]. In this study we used a clustering algorithm based on partition where the basic idea is to update the center of data points as the center of the corresponding cluster by iterative computation until the convergence is met.

K-means algorithm uses data point i in the cluster C_i ($i \in C_i$) follows (1) where $a(i)$ is the average distance between data point i and all data points in the same cluster (C_i).

$$a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j) \quad (1)$$

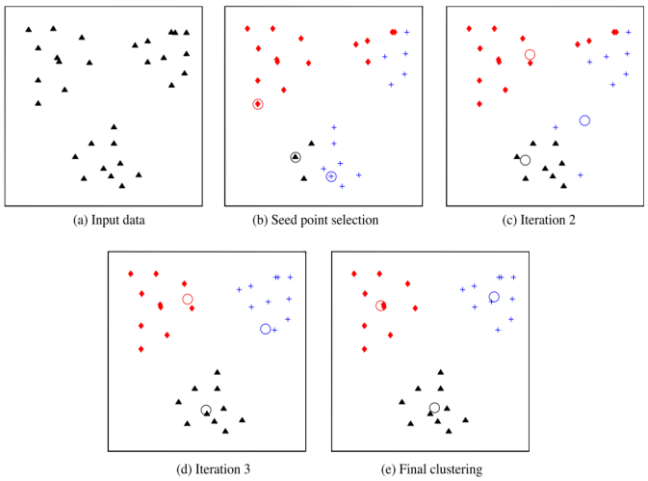

Fig. 3. Illustration of K-means algorithm.

Furthermore, $b(i)$ is the minimum average distance between data point i to all data points in other clusters (C_k) where ($C_k \neq C_i$) as shown in (2).

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \quad (2)$$

Smaller average distance of data point i to all points in any other cluster indicates the data point i is not a member of the cluster. The cluster with this small mean dissimilarity or neighboring cluster of data points i.

C. Silhoutte Coefficient

One of evaluation indicators is silhouette coefficient (SC) which evaluates the cluster based on the measure of average distance between one data point and other data points in the same clusters (cohesion) and average distance among different cluster (separation) [19]. The advantage of SC is only dependent on the partition of the data set but not on the clustering algorithm itself.

The silhouette value of single data point i shows in (3) which indicate how tightly grouped of data point in that cluster.

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad (3)$$

The Silhouette value $s(i)$ can vary between -1 and 1 can be expressed in (4) but when $s(i)$ is about 0 then $a(i)$ and $b(i)$ are approximately equal value.

$$1 \leq s(i) \leq 1 \quad (4)$$

Kaufman et al introduced the SC as shows in (5) for the optimal value of the average $s(i)$ over all data point in dataset with specific number of cluster k.

$$SC = \max_k s(k) \quad (5)$$

The monthly minimum electrical peak load dataset of Jawa Bali system from 2014 to 2020 shows in Fig. 4.

D. Dataset

This paper focuses on dataset analysis from minimum monthly Jawa Bali peak load that occur almost six years from 2014 to 2020 as shown in Fig. 4. Selection of minimum peak load than maximum and average is just because to classify the lowest peak load in a special case that may occur in this system. Low peak load condition in this system is related to national holiday and special cases such as Covid-19 pandemic as shown in Table I.

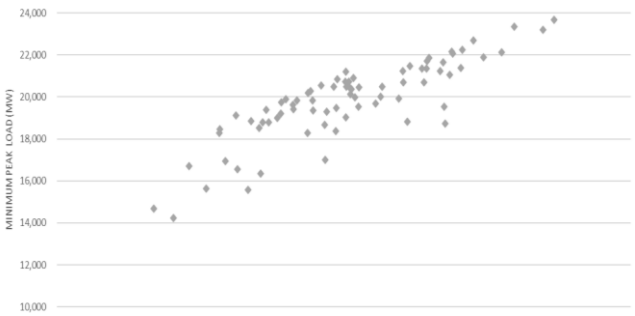

Fig. 4. Monthly minimum electrical peak load dataset.

TABLE I. LIST OF NATIONAL HOLIDAY AND EVENT IN INDONESIA

Date	Holiday and Event Name
from 10th April 2020	Indonesia large-scale social restrictions (PSBB)
from 16 th March 2020	Work from home (WFH)
1st January 2014 to 2020	New Year's Day
3-4 th June 2019 14-15th June 2018 24-25 th June 2017 6-7 th July 2016 17-18th July 2015 28-29 th July 2014	Eid al-Fitr (Idul Fitri)

III. CASE STUDIES

At 2020, Jawa Bali system peak load constantly decreases month by month in linear function as expressed in Fig. 5. The linear function is as follows f(x)=-10.444x+25114. The lower peak load started in March caused by significant decrease consumption in the industrial sector driven by WFH recommendation. The peak load in April 2020 (red line) as shown in Fig. 6 is relatively low compared to peak load profile than some year before.

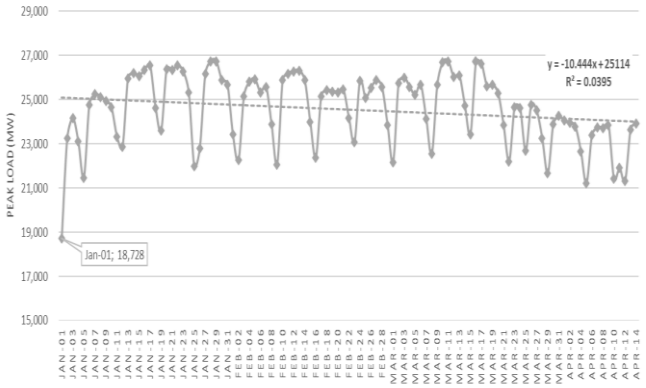

Fig. 5. Daily electrical peak load during January to April 2020.

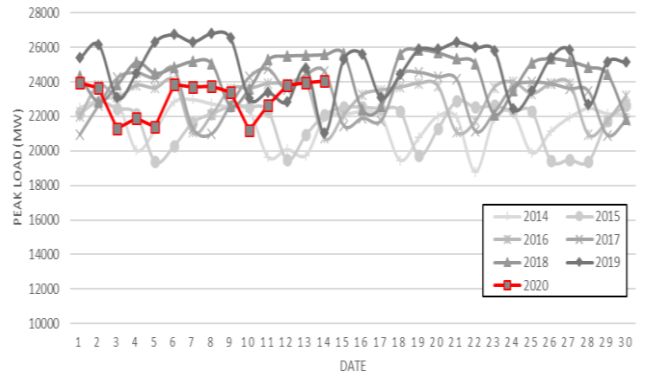

Fig. 6. April daily electrical peak load from 2014 to 2020.

The extreme minimum peak load happen in January and June or July as shown in Fig. 7. As mentioned before in Table I it caused by national holiday and event in Indonesia. This paper propose a method using K-means algorithm to analyse this low peak load phenomenon into cluster. This algorithm aims to classify peak load characteristicss especially in Covid-19 pandemic condition and national holiday. The simulation purpose is to make identification of the low peak load with similar profile and characteristic.

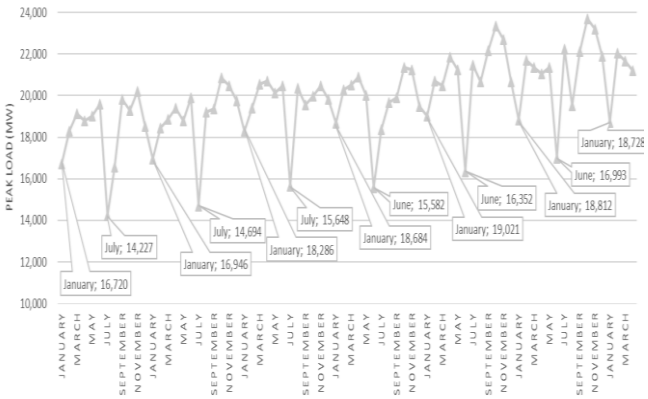

Fig. 7. Monthly minimum electrical peak load.

IV. SIMULATION RESULTS

From Table II, the electrical peak loads are optimum classified into three clusters. Simulation result of clustering shown in Fig. 8. There are three clusters where Cluster-1 (blue), Cluster-2 (red), and Cluster-3 (green). The attributes of each cluster are specified by load level (high, intermediate, and low). It can be seen in Fig. 7 and Fig. 8 that centroids during January, June, and July are relatively lower than other centroid clusters. This can be explained that in Indonesia the electricity demand in these months are in holiday (new year and Eid al-Fitr/Idul Fitri) as explained in Table I.

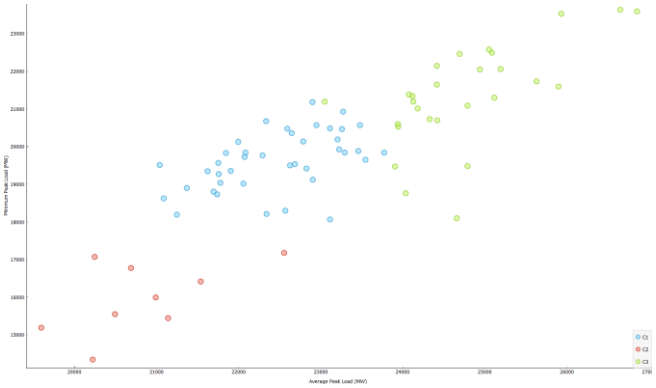

Fig. 8. Clustering result for monthly minimum electrical peak load.

TABLE II. SILHOUETTE SCORES OF CLUSTER

Number of Cluster	Silhouette
2	0.322
3	0.343
4	0.243
5	0.248
6	0.231
7	0.186
8	0.226
9	0.221
10	0.215

Simulation result from K-means clustering algorithm shown in the cluster-2 (C2) in Fig. 8 is low load level dominated by similarity load condition in special cases such as holiday and derivative impact (28-29th July and 1st August 2014) as shown in Table III.

The three high deviation peak load are in January, June and July as shown in Fig. 9 and this is match with clustering method in Fig. 8.

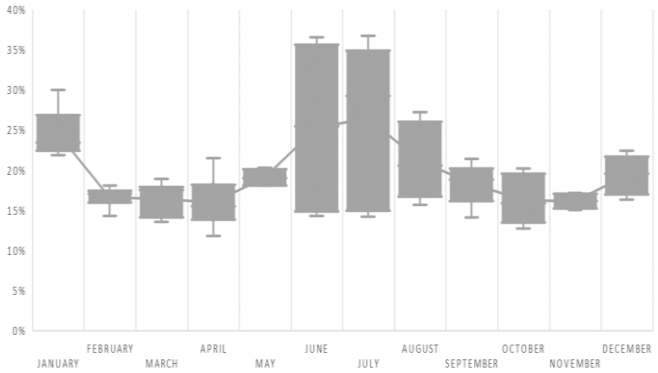

Fig. 9. Monthly electrical peak load deviation from 2014 to 2020.

TABLE III. SELECTED DATASETS IN PEAK LOAD CLUSTER

Cluster	Silhouette	Month	Year	Minimum Peak Load (MW)
C2	0.525601	January	2014	16,720
C2	0.638501	July	2014	14,227
C2	0.556393	August	2014	16,548
C2	0.519807	January	2015	16,946
C2	0.641217	July	2015	14,694
C2	0.638361	July	2016	15,648
C2	0.632389	June	2017	15,582
C2	0.609852	June	2018	16,352
C3	0.642073	March	2020	21,658
C3	0.521958	April	2020	21,214

The Covid-19 pandemic impact is classify in cluster-3 where indicate in the high load level. It means in the peak load viewpoint the Covid-19 pandemic impact cause the decreasing peak load but in case is not as low with Indonesia national holiday condition. But, in the future the system operator must prepare the lower peak load condition in next Eid al-Fitr/Idul Fitri affected by Covid-19 pandemic.

V. CONCLUSIONS

This paper present the peak load analysis from historical dataset by clustering algorithm. The clustering method used the K-means algorithm and evaluated by silhouette scores.The simulation results show there are three main clustering with similarity profiles which are low, intermediate, and high peak load level evaluated by optimum silhouette scores. We found in low peak load level classify by Indonesia national holiday with high deviation between maximum and minimum peak load in January, June, and July.

REFERENCES

[1] J. Hartono, P. A. A. Pramana, H. B. Tambunan, and B. S. Munir, 'Disturbance Magnitude Estimation using Artificial Neural Network Method', in 2019 International Conference on Electrical Engineering and Informatics (ICEEI), 2019, pp. 570–573.

[2] B. B. S. D. A. Harsono, B. S. Munir, and N. W. Priambodo, 'Lightning data mapping of West Java province', in 2017 International Conference on Electrical Engineering and Computer Science (ICECOS), 2017, pp. 300–304.

[3] J. Hartono, N. Hariyanto, F. S. Rahman, T. Kerdphol, M. Watanabe, and Y. Mitani, 'Power System Stabilizer Tuning to Enhance Kalimantan Selatan - Tengah and Kalimantan Timur System Interconnection Stability Using Particle Swarm Optimization', in 2018 5th International Conference on Electric Power and Energy Conversion Systems (EPECS), 2018, pp. 1–6.

[4] A. S. Surya, P. Awater, M. P. Marbun, and N. Hariyanto, 'Optimal Allocation of Photovoltaic in the Hybrid Power System using Knapsack Dynamic Programming', in 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), 2019, pp. 1828–1833.

[5] D. Xu and Y. Tian, 'A Comprehensive Survey of Clustering Algorithms', Ann. Data Sci., no. May, 2015.

[6] J. Xie, R. Girshick, A. Farhadi, A. L. I. Cs, and W. Edu, 'Unsupervised Deep Embedding for Clustering Analysis', vol. 48, 2016.

[7] E. Elhamifar and R. Vidal, 'Sparse Subspace Clustering: Algorithm, Theory, and Applications', IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 11, pp. 2765–2781, 2013.

[8] A. K. Jain, 'Data clustering: 50 years beyond K-means', Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[9] S. Na, L. Xumin, and G. Yong, 'Research on k-means Clustering Algorithm', 2010 Third Int. Symp. Intell. Inf. Technol. Secur. Informatics, pp. 63–67, 2010.

[10] T. M. Kodinariya, 'Review on determining number of Cluster in K-Means Clustering', Int. J. Adv. Res. inComputer Sci. Manag. Stud., vol. 1, no. 6, pp. 90–95, 2013.

[11] H. Zhou and J. Gao, 'Automatic method for determining cluster number based on silhouette coefficient', Adv. Mater. Res., vol. 951, pp. 227–230, 2014.

[12] T. Thinsungnoen, N. Kaoungku, P. Durongdumronchai, K. Kerdprasop, and N. Kerdprasop, 'The clustering validity with silhouette and sum of squared errors', in Proceedings of the 3rd

International Conference on Industrial Application Engineering, 2015, pp. 44–51.

[13] A. Starczewski, 'Performance Evaluation of the Silhouette Index', Int. Conf. Artif. Intell. Soft Comput., vol. 1, pp. 49–50, 2015.

[14] G. Chicco, 'Overview and performance assessment of the clustering methods for electrical load pattern grouping', Energy, vol. 42, no. 1, pp. 68–80, 2012.

[15] W. Labeeuw and G. Deconinck, 'Residential Electrical Load Model based on Mixture Model Clustering and Markov Models', IEEE Trans. Ind. Informatics, vol. 9, no. 3, pp. 1–9, 2013.

[16] Y. Wang, Q. Chen, and C. Kang, 'Clustering of Electricity Consumption Behavior Dynamics toward Big Data Applications', IEEE Trans. Smart Grid, vol. 7, no. 5, pp. 1–11, 2016.

[17] D. D. Sharma and S. N. Singh, 'Electrical Load Profile Analysis and Peak Load Assessment using Clustering Technique', 2014 IEEE PES Gen. Meet., 2014.

[18] WHO, 'Coronavirus Disease 2019 (COVID-19)', 2020.

[19] L. Kaufman, Finding groups in data : An introduction to cluster analysis. Hoboken: Wiley-Interscience, 1990.