Revisiting Fuel Subsidies in Indonesia using *K*-Means, PAM, and CLARA

Fajar Agung Prasetyo[†], Rezzy Eko Caraka[*†] *Member IAENG*, Yunho Kim[*], Noor Ell Goldameir, Sulistyowati, Avia Enggar Tyasti, Prana Ugiana Gio, Faisal Anggoro, Muthia Ramadhani, Bens Pardamean

Abstract—**Indonesia is one of the countries in the world that still applies subsidies for fuel oil. By the law, the Indonesian government must ensure the supply and distribution of fuel for all Indonesian people. To implement this policy properly, understanding the pattern of fuel consumption is fundamental. In this study, clustering will be used to determine the categories of districts and cities based on subsidized fuel consumption patterns. Our research used several methods to compare which cluster method is the most optimal; the methods include *k*-means, Partitioning Around Medoids (PAM), and Clustering Large Applications (CLARA). The results show that *k*-means is the best clustering method with the highest Dunn index and Silhouette coefficient compared to others. The optimum cluster number we get is two and represents fuel consumption from several districts and cities. Some districts and cities had under-average consumption and needed to be monitored.**

Index Terms— **CLARA, fuel subsidies, *k*-means, PAM**

I. Introduction

Current fuel consumption increases, especially in Indonesia [1]–[5]. Every year the government sets a quota for subsidized fuel to ensure the people's need for power. However, the government has tried to reduce fuel consumption in Indonesia because it wants to switch to more environmentally friendly fuels such as electricity [6]–[9]. This is done gradually by mixing fuel, especially diesel fuel, with fatty acid methyl ether (fame). Fame is a fluid usually obtained from plant oils that can be used to keep the cetane number in diesel fuel [10]. The most widely used subsidized fuel type is diesel fuel [1], [11]. Because it is widely used, the government is trying to reduce pure diesel for energy. The use of fame has been determined through established regulations, namely, mixing starting at 20% or B20 and will continue to increase. It has reached 30% of its fame and is called B30.

Government policies related to the use of subsidized fuel are often not well targeted because they are considered to be still usable by the user sector, who should not use subsidized fuel. Supervision of subsidized energy needs to be increased. In line with the Indonesian government's policy of limiting the use of fossil energy and gradually shifting to environmentally friendly fuels, the determination of the subsidized fuel quota is considered to support this policy [12]. Determining the right amount of sponsored fuel quota can limit fuel use. However, the government needs to pay attention that the distribution of fuel in Indonesia is still not evenly distributed. So it is necessary to create a category of cities and districts to determine which ones need to be prioritized to get subsidized fuel.

Data on fuel consumption is fundamental in supporting the government's determination to take policies. Through this data, the government can efficiently conduct the analysis, but on the other hand, it is necessary to have completeness and proper interpretation of the data [13]. In this study, we use data on the realization of subsidized fuel for the type of diesel fuel from 2016 to June 2021. The data obtained is in the form of completion per month for each district and city in Indonesia. The research only focused on diesel fuel types because regulations related to subsidized diesel fuel tend to be more stable and have not experienced significant changes in recent years. Meanwhile, this situation is very different from the type of kerosene. The regulation has an unstable impact on data availability due to the conversion of kerosene to LPG gas which is still ongoing today.

Based on the data used, it can be seen that the time-series dataset is an appropriate data form to facilitate data analysis because it had historical information [14]–[17], about the consumption of fuel for each district and city. The time series form will help determine patterns and behavior [18]–[20] for each city and district's consumption in using subsidized fuel. To help facilitate the analysis of the pattern and behavior of fuel consumption, this study proposes a clustering technique to categorize districts and cities with similar designs and behavior of subsidized fuel consumption. The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis

Manuscript received January 01 2022; revised July 09, 2023. Rezzy Eko Caraka is partially supported by the National Research Foundation of Korea (NRF-2023R1A2C1006845). Yunho Kim acknowledges support from the National Research Foundation of Korea (NRF-2022R1A5A1033624, NRF-2023R1A2C1006845).[†]Fajar Agung Prasetyo and Rezzy Eko Caraka equal as the first authorship.

F.A.Prasetyo is a postgraduate student, Faculty of Economics and Business, Universitas Indonesia, Indonesia. (e-mail: fajar.agung02@ui.ac.id)

R.E.Caraka is an Associate Researcher, Research Center for Data and Information Sciences, National Research and Innovation Agency (BRIN), Indonesia. (Corresponding author e-mail: rezzy.eko.caraka@brin.go.id).

Y.Kim is an Associate Professor, Department of Mathematics, Ulsan National Institute of Science and Technology. South Korea, (Corresponding author e-mail: yunhokim@unist.ac.kr)

N.E.Goldameir is a Senior Lecturer, Department of Mathematics, Statistics Study Program, Riau University, Pekanbaru, Riau, Indonesia. (e-mail: noorellgoldameir@gmail.com)

Sulistyowati is a Senior Lecturer, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia. (e-mail: aryis77@gmail.com)

A.E.Tyasti is a Senior Lecturer, International Trade of ASEAN and RRT Region, Polytechnic of APP, DKI Jakarta, Indonesia. (e-mail: aviaenggar0210@gmail.com)

P.U.Gio is a Senior Lecturer, Department of Mathematics, Universitas Sumatera Utara, Medan 20155, Indonesia. (e-mail: prana@usu.ac.id)

F.Anggoro is a postgraduate student, Faculty of Economics and Business, Universitas Indonesia, Indonesia. (e-mail: faisal.anggoro@ui.ac.id)

M.Ramadhani is a postgraduate student, Faculty of Economics and Business, Universitas Indonesia, Indonesia. (e-mail: muthia.ramadhani@ui.ac.id)

B.Pardamean is a Professor, Computer Science Department, BINUS Graduate Program - Master of Computer Science Program, Bina Nusantara University, Jakarta, Indonesia. (e-mail: bpardamean@binus.edu)

[21]–[23]. Clustering algorithms can be implemented in statistics, computer science, and machine learning [24]–[30]. Since the method clusters using extracted global measures, it reduces the dimensionality of the time series and is much less sensitive to missing or noisy data [31]–[35].

Nowadays, the clustering technique is used in several cases that need to be done, like image segmentation, object recognition, and information retrieval. The application of clustering in agriculture is beneficial for harvesting with various techniques that have been investigated using different types of sensors and their combinations and other image processing techniques [36]. Clustering is the first step in implementing deep learning and could be the vehicle for translating big biomedical data into improved human health [37]. Medical data are complex, high-dimensional, and heterogeneous; biomedical data remains a key challenge in transforming health care.

This study tried to compare several clustering methods to determine which plan has the best results in clustering based on the data set. Each technique will be applied to obtain detailed and comprehensive results, including the *k*-means, PAM, and CLARA. This method is used to determine the similarity of consumption patterns of subsidized fuel in each district and city in Indonesia and to describe consumption patterns based on historical data on subsidized fuel consumption. This study had an essential role because of the unequal distribution of fuel, and there are still many distributions of subsidized energy that are not yet on target. Besides that, the government also needs to know the priority areas for receiving subsidized fuel so that subsidized fuel can be controlled.

II. Literature review

Determining the amount of the subsidy quota is fundamental to ensuring the use of fuel is right on target. Currently, there are no studies that focus on discussing fuel distribution policies that use big data as a basis for decision-making in Indonesia. Several existing studies related to fuel in Indonesia mostly discuss the impact of fuel consumption on the environment and even on economic growth and the factors that influence it. Some studies in big data are usually related to the smart city if we are concerned about the government environment. The relationship between fossil fuel consumption, carbon dioxide emissions, and economic growth, the impacts of energy conservation, as well as the projection of energy mix in Indonesia [6], [38]. Both studies propose the same recommendations that can be given to the Indonesian government specifically improving the provision of a more environmentally friendly non-fossil energy source and increasing awareness of environmental health and developing sustainable industries based on renewable energy. Based on the report of The Ministry of Energy and Mineral Resources and Ministry of Finance, fuel subsidies, if targeted well, can help reduce the burden of international oil prices on the poor; on the other hand, the budgets for such subsidies tend to increase over time, ultimately putting pressure on fiscal capacity. Unfortunately, the elimination of subsidies can not be done easily because the government has the mandate to provide energy for the poor and there are political issues. So the best recommendation, for now, is to make fuel subsidies more targeted.

The current technological developments can assist the government to determine policies that are in accordance. One of them is Big Data which has been widely used in various fields, including the government sector. According to [39], there are some basic criteria that the government should have to use big data such as the quality standards of data, standards regarding privacy and data protection, policy regarding the ownership of unstructured and structured data and standards regarding safety and protection of storing of data. Big Data can foster collaboration; create real-time solutions to challenges in several sectors and usher in a new era of policy and decision-making [13].

One of the most widely used big data analysis methods is clustering. The clustering method also has been widely used in various research fields. Research in [40] describes the role of clustering in determining strategies for finding and analyzing the periodicity of energy in time series that can be deepened in detail and extracts information at various levels in detail.

A. *K-means*

In simple terms, the *k*-means algorithm consists of three parts, the first was to initialize the initial cluster center or what is often known as the centroid [41]. This initiation can be done using random or heuristic algorithms or by iterating so that convergence occurs. The iteration itself consists of two parts, the first by determining the group affiliation from each point to the nearest centroid [42]. Next, the centroid position was updated based on the average value of the group member's field [43]. As [44] found that the *k*-means method was considered vulnerable to data outliers and noise, this statement was supported by [45] who got the same result from the research. One of the clustering methods that are considered capable of handling data outliers and noise is Partitioning Around Medoid.

At first, the data was in one group. In this group, the centroid initialization was determined, building group affiliation from each point in each centroid. Each dot that is a group member's notation will update its membership based on cooperation to the nearest centroid. In line with this, the process will be run until there is no change in membership at all points. Therefore, this condition was convergent [46].

Pseudo-code *k*-means
Input: $D=d_1, d_2,d_3,d_4,\ldots,d_n$ k= number of clusters
Output: set k clusters $C=c_1,c_2,\ldots,c_k$
Method : **Step 1** Choose k variables randomly from a set of n observations as the centroid. **Step 2** Calculate the similarity between the centroid and non-centroid observations. **Step 3** Get the cluster members by selecting the variable that has the most remarkable similarity between non-centroid and centroid observations **Step 4** Update the centroid by calculating the average of the cluster members **Step 5** Repeat **steps 2 to 5** until there is no change in the cluster members.

B. *Partitioning Around Medoids (PAM)*

PAM was a type of *k*-medoids clustering technique. PAM was used to overcome the sensitivity to outliers that often occurs when the *k*-means method was used [44]. Based on [47], PAM randomly partitions the data set into several subgroups and then increases the medoid cluster iteratively to minimize the objective function so that it was considered one of the clustering methods that can handle noise and outliers. However, according to [48], the PAM method was not suitable for use on large datasets because it has low scalability. In the *k*-medoids approach, actual data represents the cluster, unlike *k*-means which uses the average value as the centroid. In PAM, it is necessary to determine the number of clusters to be built in advance. The first thing that needs to be done is select an object from the data set as a representative (seed). This activity continues until each object becomes part of the cluster based on its proximity to the representative object. For each representative object (O_i), an object (O_{random}) was randomly selected to compare its cost value. Suppose the cost(s) for exchanging Oi representative with O_{random} is less than 0. In that case, O_i position will be replaced by O_{random} as the representative and will continue until there was no change.

C. Clustering Large Application (CLARA)

CLARA was developed by Kaufman and Rousseeuw in 1990. This method generates *k*-medoids to overcome many objects in the data to be clustered to reduce data processing time [49]. CLARA was a development of the PAM method which is intended to handle clustering for very large amounts of data. CLARA has high scalability but low sensitivity to outliers and noise [19]. The quality of the final medoids was measured by the average dissimilarity of each object in the data set. In addition, each cluster's medoids are defined as a cost function. The first step in applying CLARA was to randomly determine the number of samples from the existing data set with a fixed size. Next, choose a representative object or medoid and observe all current samples against the closest medoid. Calculate the mean or the number of dissimilarities from observations to the nearest medoid to obtain a good cluster quality. The last step was to call the sample with the lowest mean value and repeat the previous steps until the last cluster was obtained.

III. Methodology

This study used data reports on the distribution of fuel in all regions of Indonesia conducted by two companies that have the authority to distribute subsidized fuel (Pertamina and AKR Corporindo). These two reports will be combined into new raw data with a uniform format to facilitate further processing. The data is described in a time series for each district and city in Indonesia with dimensions: name of the province, city, consumption of fuel, longitude, and latitude. And we use R studio to process data and run on an Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz 2.60 GHz with 8 GB of RAM. The period of the data used in this study was data from January 2016 to June 2021. The data used in this study is only limited to data on the consumption of subsidized fuel for diesel oil, not including other types of fuel. The name and number of districts used refers to the Ministry of Home Affairs and are used as a naming standard in the preparation of raw data.

Data cleansing is carried out to select the data needed for analysis. Cleansing data is done by removing duplicate, empty, or negative data that has been confirmed to the data source but considered to have no meaning in the fuel consumption data [50]. Based on 132 monthly reports that data cleansing has been carried out, 514 districts or cities have consumption of subsidized fuel. Furthermore, to determine the number of clusters, this study uses the Silhouette method, according to [51], the Silhouette method can provide recommendations for the optimal number of clusters for large datasets. The optimal number of clusters for each method can be seen in Table 1.

TABLE 1
OPTIMAL NUMBER OF CLUSTER

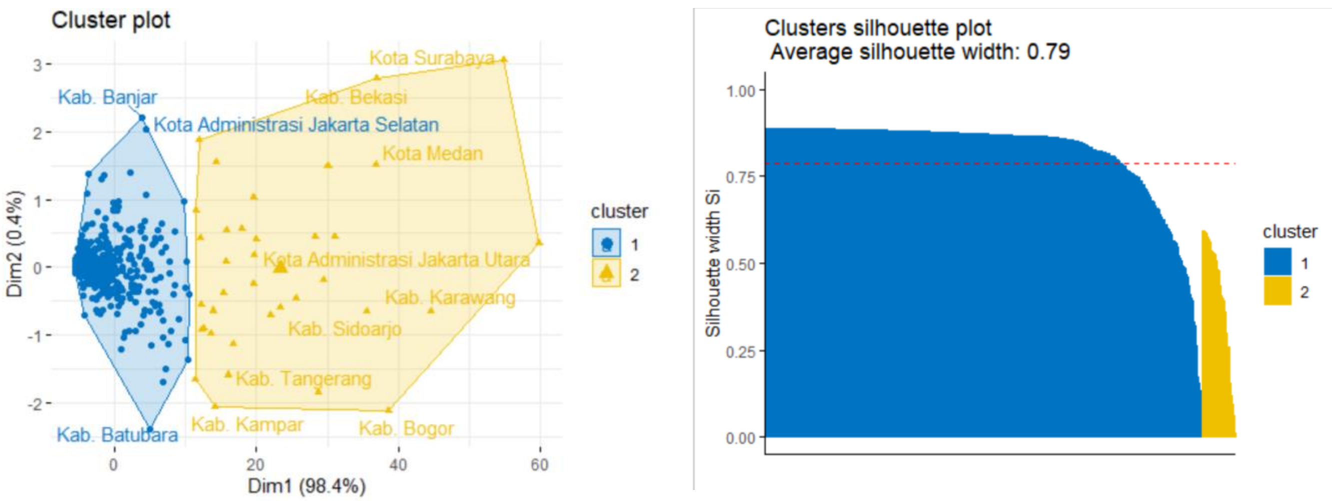
Clustering Method	Optimal Number of Cluster
k-means	2
Partitioning Around Medoids	2
Clustering Large Applications	2

IV. Data Analysis and findings

Clustering was grouping the objects in the data into two or more groups based on the characteristics between objects [48]. There are various kinds of clustering techniques, but in these studies we used time-series data. Objects in data clustering have different distribution patterns, so the grouping of time series data will provide an overview based on the observed time [31], [40]. After preprocessing the data described previously, the next step was to study the subsidized fuel consumption data periodically or monthly. Each city or district has monthly fuel realization data from January 2016 to June 2021.

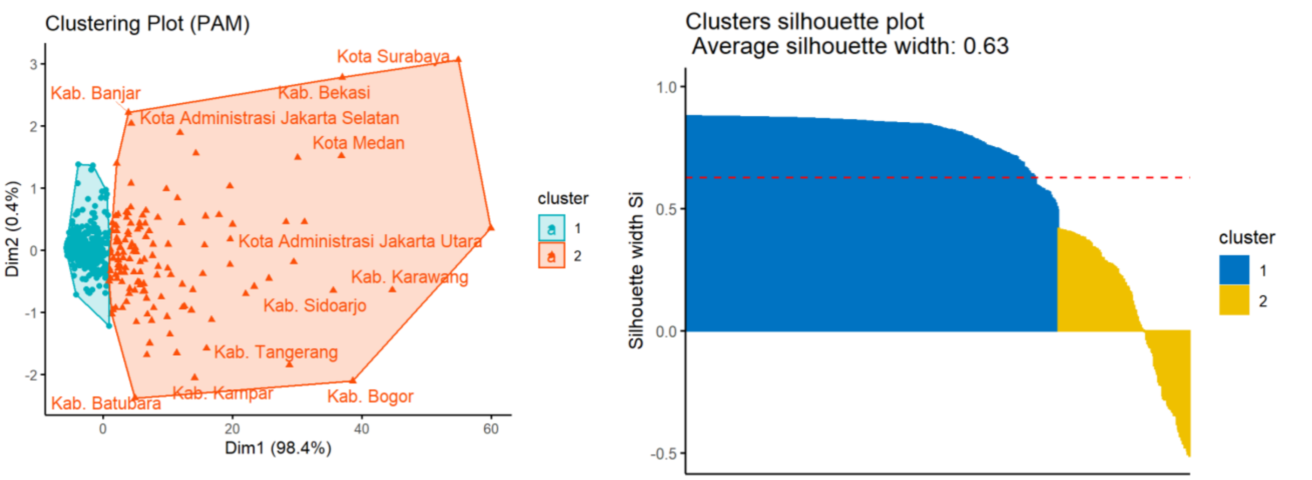
This historical data will be clustered using the previously mentioned techniques, namely *k*-means, PAM, and CLARA. Before that, to get an ideal number of clusters, a calculation will be made using the silhouette method. The results of the silhouette method obtained that the optimum number of clusters for each clustering technique was the same, with 2 clusters. The first clustering technique used is *k*-means with 2 clusters as the result of the optimum cluster. In this technique, the first thing to do is to determine the seed set and then be processed to select the category of a particular object cluster [50]. The results of the *k*-means for subsidized fuel consumption data can be seen in Fig. 1a and Fig. 1b with an average silhouette value of 0.79. The second clustering technique used is PAM, from calculating the optimum number of clusters using silhouette obtained 2 clusters as the optimum number of clusters represented in Fig. 2a. The clustering results show that the average silhouette value obtained for these two clusters is 0.63, as shown in Fig. 2b.

In the following clustering technique, we used CLARA with 2 clusters and produced an average silhouette value of 0.65, which represents in Fig. 3a and Fig. 3b. The cluster results with the CLARA technique are almost the same as the cluster results with the PAM technique because CLARA is an extension of the PAM clustering method for large data sets and reduces computation time for extensive data [20]. The results of comparing the clustering results were validated using Connectivity, Dunn, and Silhouette with the results as shown in Table 1. Based on the comparison data, the *k*-means clustering technique with cluster 2 has the highest connectivity and silhouette values compared to other methods. The Dunn *k*-means value has the highest points, but for the number of clusters, there are six clusters [51]. Based on the results of this analysis.



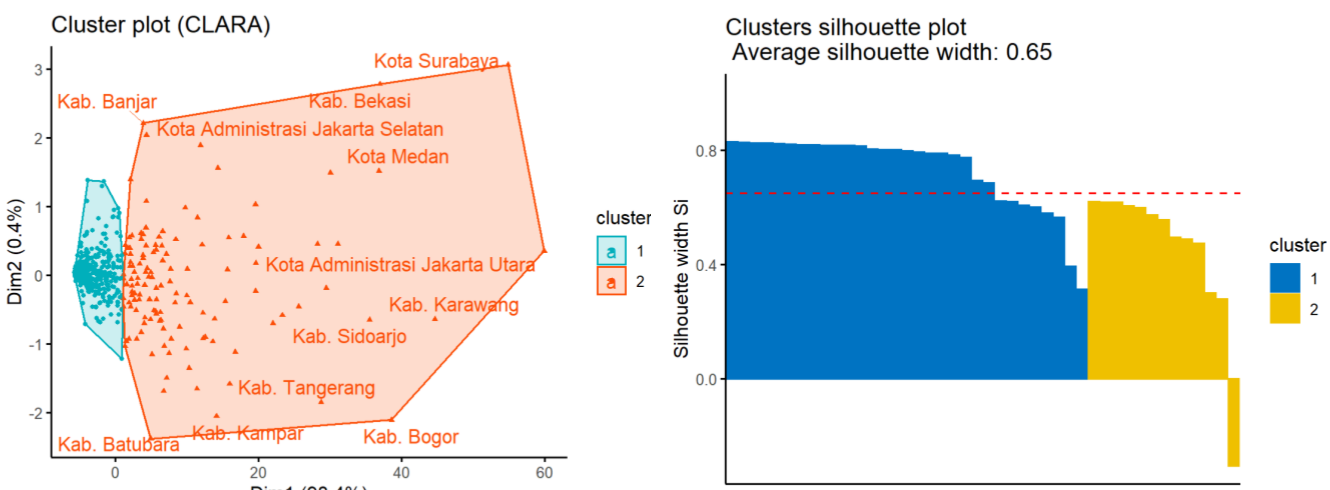
a b

Fig. 1. *k*-means Clustering Towards Fuel Consumption (a), and Clusters Silhouette (b)



a b

Fig. 2. PAM Clustering Towards Fuel Consumption (a), and Clusters Silhouette (b)



b

a

Fig. 3. CLARA Clustering Towards Fuel Consumption (a), and Clusters Silhouette (b)

Fig. 4 shows the spatial clustering using *k*-means to represent the vulnerability of fuel consumption in Indonesia. To create a spatial cluster, we use the Euclidean distance, let $s \subseteq \mathbb{R}^d$ with time series observations $y_1, y_2, \dots, y_n$ for any finite $C \subseteq \mathbb{R}^d$, we define $dist^2(y,C) = \min_{c\epsilon S} dist^2\|y,C\|^2$. Then, to find the weighting function $w(y)$ for each $y \in S$, The goal of weighted *k*-means clustering is to find a set of clusters $C \subset \mathbb{R}^d$ and minimize $cost_w(S,C)$ with the following criteria

$$cost_w(S,C) = \sum_{y \in S} w(s).dist^2(y,C)$$

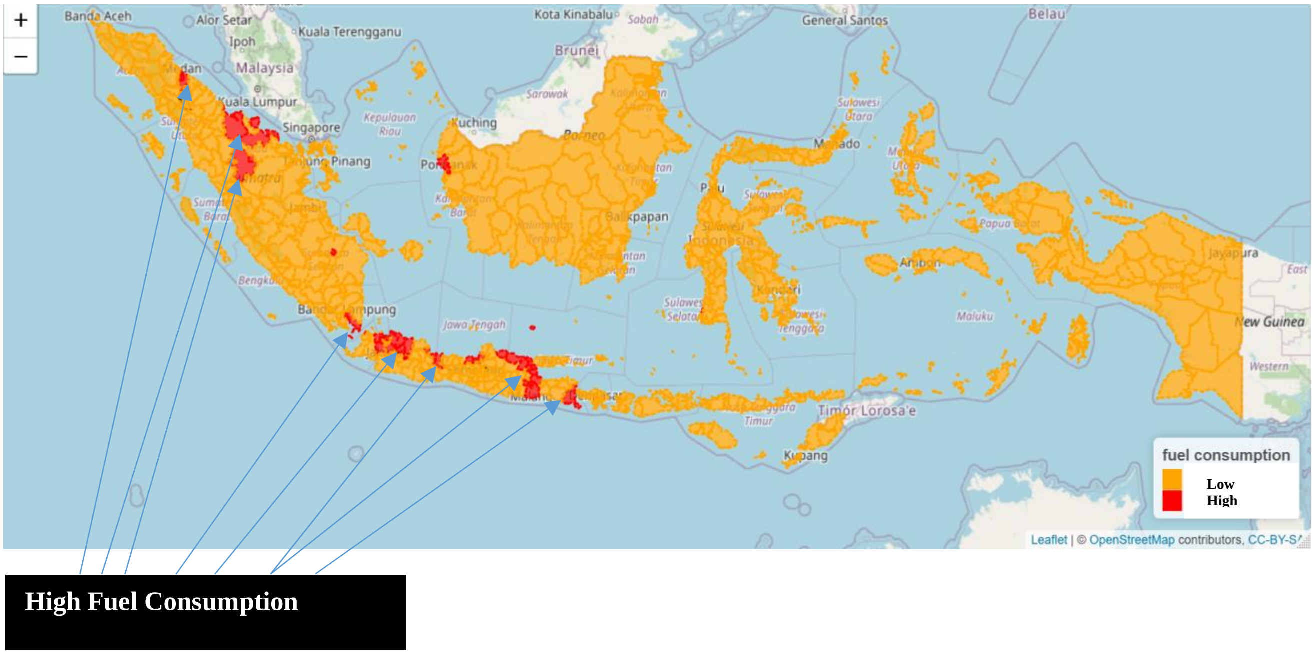


Fig. 4. Spatial Clustering Towards Fuel Consumption using *K*-Means

TABLE 2
CLUSTERING MEASUREMENT

Methods	Validity measures	Number of Clusters 2	3	4	5
K-Means	Connectivity	7.327	22.003	27.479	34.4774
	Dunn	0.0434	0.0188	0.0386	0.0312
	Silhouette	0.7884	0.6711	0.6523	0.6049
PAM	Connectivity	22.3377	32.7571	34.1587	50.6056
	Dunn	0.0093	0.0067	0.0081	0.0084
	Silhouette	0.6278	0.5561	0.5421	0.5029
CLARA	Connectivity	23.1234	25.2004	41.5853	47.7361
	Dunn	0.0086	0.0068	0.0094	0.0123
	Silhouette	0.6349	0.5534	0.5621	0.5566

Table 1 represents the clustering evaluation. These urban districts have similarities in terms of subsidized fuel consumption which is very high compared to other urban communities in Indonesia. The trend of consumption from year to year also continues to increase. The dark color on the map is a cluster of urban districts with fuel consumption above the national average. It indicates that subsidized fuel consumption in Indonesia is generally still in the normal range to some extent. However, it should be noted that there are still several urban districts that have very low consumption levels, so the government needs to carry out more supervision for these areas [50]. The method built in this research will help address the breakdown of the distribution of fuel distribution in Indonesia and evaluate appropriately subsidized fuel. However, we need to pay attention to the consumption patterns of each region in Indonesia for certain months because there are religious celebrations that have quite an effect on consumption patterns.

To compare the mapping results from clustering, we also plotted the map for the median results from each district city based on data from the beginning of the month to the end of the month in the data set. We use the limit of fuel consumption above 10,000 KL per month classified as high consumption while below 10,000 KLper months including low consumption. The mapping results can be seen in Fig. 5. It can be obtained that the spatial clustering based on the median value does not have a significant difference when compared to the spatial time series based on the data from clustering with *k*-means.

In this analysis, two variables are used to assess the relationship of fuel consumption, namely the number of residents and the fuel distribution network in that area. Cluster 1 has a strong association with regencies and cities with high fuel consumption, while cluster 2 has medium to low consumption. The top ten for cluster 1 as shown in Table 3 are mostly big cities in Indonesia and mostly located on the island of Java due to the adequate fuel distribution network (facilities and infrastructure) and a large number of people in the area. Based on data from the Ministry of Internal Affairs, the population for each of these areas reaches more than 2 million people.

Bottom ten cluster 2 is dominated by districts in Papua as shown in table 4. This is considered reasonable because currently the fuel distribution facilities and facilities are still limited. In addition, the population in Papua is also not as much as the population in other regions. The anomaly in cluster 2 is the existence of the Kepulauan Seribu which is included in this cluster. Kepulauan Seribu is one of the regencies in DKI Jakarta as a state capital. Based on data from the Central Bureau of Statistics, the population in the Kepulauan Seribu is far less than in other areas in DKI Jakarta. In addition, there are almost no cars so the use of diesel oil is very little because most of the users of diesel fuel are cars.

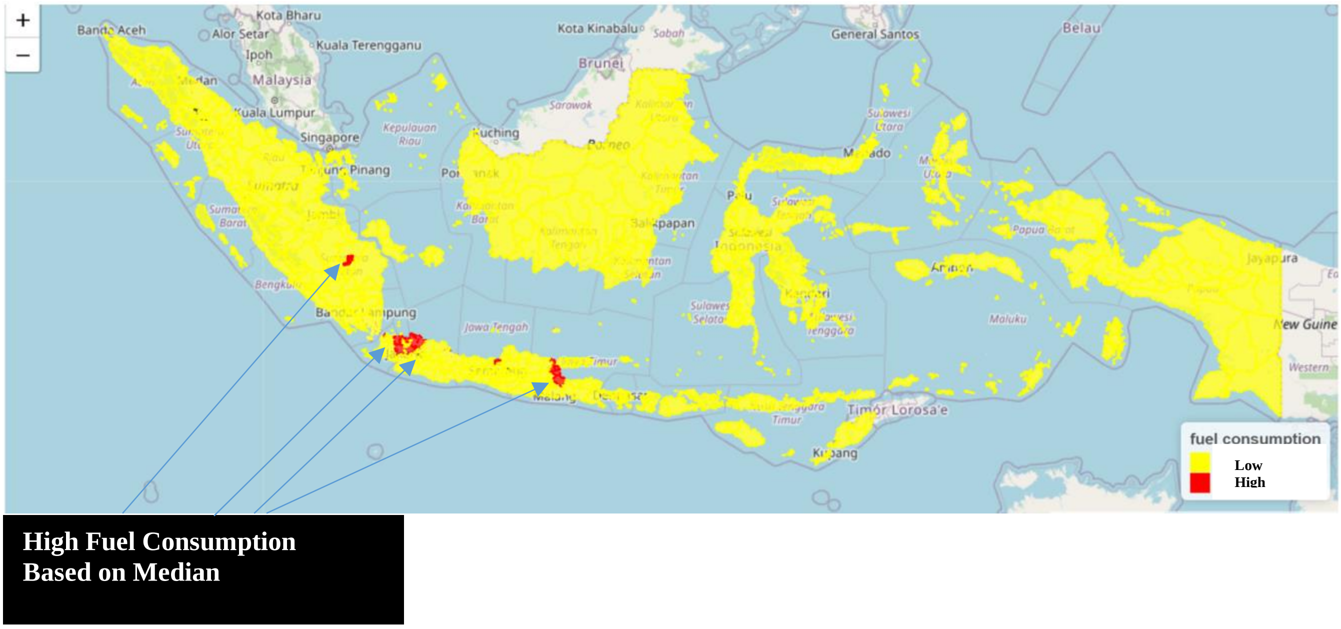


Fig. 5. Spatial Clustering Towards Fuel Consumption using Median

TABLE 3
TOP TEN CLUSTER 1

No.	Area	Province
1	Kota Administrasi Jakarta Utara	DKI Jakarta
2	Kota Surabaya	Jawa Timur
3	Kab. Karawang	Jawa Barat
4	Kab. Bogor	Jawa Barat
5	Kota Medan	Sumatera Utara
6	Kab. Bekasi	Jawa Barat
7	Kab. Sidoarjo	Jawa Timur
8	Kota Cilegon	Banten
9	Kota Administrasi Jakarta Timur	DKI Jakarta
10	Kota Tangerang	Banten

TABLE 4
BOTTOM TEN CLUSTER 2

No	Area	Province
1	Kab. Nduga	Papua
2	Kab. Puncak	Papua
3	Kab. Administrasi Kepulauan Seribu	DKI Jakarta
4	Kab. Mamberamo Tengah	Papua
5	Kab. Intan Jaya	Papua
6	Kab. Konawe Kepulauan	Sulawesi Tenggara
7	Kab. Deiyai	Papua
8	Kab. Lanny Jaya	Papua
9	Kab. Tana Tidung	Kalimantan Utara
10	Kab. Tolikara	Papua

V. CONCLUSION

This research applied clustering to analyze the consumption of subsidized fuel in Indonesia. The clustering technique is used to obtain consumption patterns from each district and city. It should be understood that the consumption behavior of each district was influenced by policies in that area. The framework we proposed in this study is used to analyze and find patterns of fuel consumption by using clustering techniques. Based on the result, although the dataset used in clustering is quite large it does not necessarily mean that the CLARA method is the best method for clustering large data. Validation testing was needed to determine the most suitable method.

For further research, if only focusing on the overall consumption, we may not know that there is an over-quota in the use of subsidized fuel. Even though Indonesia has more than 500 regencies and cities that have the right to use subsidized fuel. The grouping results obtained from this study can describe the distribution pattern for each region and can be used to detect problems that may arise such as the possibility of improper use of fuel.

This study was the basis of a model for the knowledge base in the development of decision-making support. In future research, researchers can make predictions about future fuel consumption. It will be very helpful in determining the amount of the fuel subsidy quota that must be provided by the government. In addition to the nature of the discrepancy in the forecast results, further research needs to consider several other factors that affect fuel consumption.

VI. PRACTICAL IMPLICATION

This research was conducted to assist the government in utilizing resources in the form of data owned by the clustering method. Utilization of subsidized fuel consumption report data in this study can assist the government in providing an overview of subsidized fuel consumption patterns in Indonesia so that the government can determine appropriate policies related to subsidized fuel by using subsidized fuel consumption report data as a basis for decision-making. Nationally subsidized fuel quotas have been determined yearly, but the challenge that often arises is ensuring the distribution of subsidized fuel is right on target. This research can provide insight for the government regarding district/city fuel consumption which has so far escaped scrutiny, such as areas with low consumption but the quota set by the government is very high. The government can determine which areas are prioritized in monitoring the distribution of subsidized fuel by referring to areas that are members of clusters with high vulnerability to the use of subsidized fuel. The high amount of subsidized fuel consumption in an area requires high supervision, considering that the number of subsidized fuel quotas is limited. This must be done to ensure balanced fuel distribution in an area. In addition, regions that are members of clusters with low vulnerability to the use of subsidized fuel can be used as a reference in reviewing the amount of the subsidized fuel quota that has been determined, especially for regions with a substantial remaining quota. A low level of fuel consumption can indicate problems in the supply and distribution of fuel so that the government can focus more on dealing with this problem. Furthermore, the results of this clustering can be used to evaluate subsidized fuel consumption during the current year so that the determination of quotas will be more accurate.

REFERENCES

[1] Akhmad and Amir, "Study of Fuel Oil Supply and Consumption in Indonesia," *International Journal of Energy Economics and Policy* |, vol. 8, no. 4, pp. 13–20, 2018, [Online]. Available: http:www.econjournals.com

[2] T. E. Bimanatya and T. Widodo, "Fossil fuels consumption, carbon emissions, and economic growth in Indonesia," *International Journal of Energy Economics and Policy*, vol. 8, no. 4, pp. 90–97, 2018.

[3] S. Mishra, K. Anand, and P. S. Mehta, "Predicting the Cetane Number of Biodiesel Fuels from Their Fatty Acid Methyl Ester Composition," *Energy and Fuels*, vol. 30, no. 12, pp. 10425–10434, 2016, doi: 10.1021/acs.energyfuels.6b01343.

[4] Akhmad and Amir, "Study of fuel oil supply and consumption in indonesia," *International Journal of Energy Economics and Policy*, vol. 8, no. 4, pp. 13–20, 2018.

[5] V. J. Schwanitz, F. Piontek, C. Bertram, and G. Luderer, "Long-term climate policy implications of phasing out fossil fuel subsidies," *Energy Policy*, vol. 67, pp. 882–894, 2014, doi: 10.1016/j.enpol.2013.12.015.

[6] T. E. Bimanatya and T. Widodo, "Fossil Fuels Consumption, Carbon Emissions, and Economic Growth in Indonesia," *International Journal of Energy Economics and Policy* |, vol. 8, no. 4, pp. 90–97, 2018, [Online]. Available: http:www.econjournals.com

[7] S. Sunaryo, Suhartono, and A. J., "Double Seasonal Recurrent Neural Networks for Forecasting Short Term Electricity Load Demand in Indonesia," in *Recurrent Neural Networks for Temporal Data Processing*, 2011, pp. 1–16. doi: 10.5772/15062.

[8] Suhartono, I. Puspitasari, M. S. Akbar, and M. H. Lee, "Two-level seasonal model based on hybrid ARIMA-ANFIS for forecasting short-term electricity load in Indonesia," in *ICSSBE 2012 - Proceedings, 2012 International Conference on Statistics in Science, Business and Engineering: "Empowering Decision Making with Statistical Sciences,"* 2012. doi: 10.1109/ICSSBE.2012.6396642.

[9] R. E. Caraka, R. C. Chen, T. Toharudin, B. Pardamean, S. A. Bakar, and H. Yasin, "Ramadhan short-term electric load: A hybrid model of cycle spinning wavelet and group method data handling (CSW-GMDH)," *IAENG Int J Comput Sci*, vol. 46, no. 4, 2019.

[10] S. Mishra, K. Anand, and P. S. Mehta, "Predicting the Cetane Number of Biodiesel Fuels from Their Fatty Acid Methyl Ester Composition," *Energy and Fuels*, vol. 30, no. 12, pp. 10425–10434, Dec. 2016, doi: 10.1021/acs.energyfuels.6b01343.

[11] G. Schwartz and B. Clements, "GOVERNMENT SUBSIDIES," 1999.

[12] V. J. Schwanitz, F. Piontek, C. Bertram, and G. Luderer, "Long-term climate policy implications of phasing out fossil fuel subsidies," *Energy Policy*, vol. 67, pp. 882–894, Apr. 2014, doi: 10.1016/j.enpol.2013.12.015.

[13] L. F. Luna-Reyes, J. Zhang, S. Mellouli, Association for Computing Machinery, and ACM Digital Library., "Big Data and e-Government: Issues, Policies, and Recommendations," *International conference on digital government research*, vol. 14, pp. 1–10, 2013.

[14] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering - A decade review," *Inf Syst*, vol. 53, pp. 16–38, May 2015, doi: 10.1016/j.is.2015.04.007.

[15] I. F. Hassan, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, "Deep learning for time series classification: a review," *Data Min Knowl Discov*, vol. 33, no. 4, pp. 917–963, Jul. 2019, doi: 10.1007/s10618-019-00619-1.

[16] H. Izakian, W. Pedrycz, and I. Jamal, "Fuzzy clustering of time series data using dynamic time warping distance," *Eng Appl Artif Intell*, vol. 39, pp. 235–244, Mar. 2015, doi: 10.1016/j.engappai.2014.12.015.

[17] R. Ma and R. Angryk, "Distance and Density Clustering for Time Series Data," in *IEEE International Conference on Data Mining Workshops, ICDMW*, IEEE Computer Society, Dec. 2017, pp. 25–32. doi: 10.1109/ICDMW.2017.11.

[18] P. Montero and J. A. Vilar, "TSclust: An R Package for Time Series Clustering," 2014. [Online]. Available: http://www.jstatsoft.org/

[19] S. K. Popat and M. Emmanuel, "Review and Comparative Study of Clustering Techniques," 2014. [Online]. Available: www.ijcsit.com

[20] A. Radovanovic, X. Ye, J. v. Milanovic, N. Milosavljevic, and R. Storchi, "Application of the k-medoids partitioning algorithm for clustering of time series data," in *IEEE PES Innovative Smart Grid Technologies Conference Europe*, IEEE Computer Society, Oct. 2020, pp. 645–649. doi: 10.1109/ISGT-Europe47291.2020.9248796.

[21] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," 2000.

[22] F. Anggoro *et al.*, "Revisiting Cluster Vulnerabilities towards Information and Communication Technologies in the Eastern Island of Indonesia Using Fuzzy C Means," *Sustainability*, vol. 14, no. 6, p. 3428, Mar. 2022, doi: 10.3390/su14063428.

[23] B. I. Nasution, R. Kurniawan, T. H. Siagian, and A. Fudholi, "Revisiting social vulnerability analysis in Indonesia: An optimized spatial fuzzy clustering approach," *International Journal of Disaster Risk Reduction*, vol. 51, no. May, p. 101801, 2020, doi: 10.1016/j.ijdrr.2020.101801.

[24] R. Xu and D. Wunsch, "Survey of Clustering Algorithms," *IEEE Trans Neural Netw*, vol. 16, no. 3, pp. 645–678, May 2005, doi: 10.1109/TNN.2005.845141.

[25] R. E. Caraka *et al.*, "Cluster Around Latent Variable for Vulnerability Towards Natural Hazards, Non-Natural Hazards, Social Hazards in West Papua," *IEEE Access*, vol. 9, pp. 1972–1986, 2021, doi: 10.1109/ACCESS.2020.3038883.

[26] J.-A. Mueller and F. Lemke, *Self-organising data mining: An intelligent approach to extract knowledge from data*. 1999.

[27] V. Nourani, A. H. Baghanam, J. Adamowski, and M. Gebremichael, "Using self-organizing maps and wavelet transforms for space-time pre-processing of satellite precipitation and runoff data in neural network based rainfall-runoff modeling," *J Hydrol (Amst)*, vol. 476, pp. 228–243, 2013, doi: 10.1016/j.jhydrol.2012.10.054.

[28] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit Lett*, 2010, doi: 10.1016/j.patrec.2009.09.011.

[29] P. Willett, "Recent trends in hierarchic document clustering: A critical review," *Inf Process Manag*, vol. 24, no. 5, pp. 577–597, 1988, doi: 10.1016/0306-4573(88)90027-1.

[30] B. Basaran and F. Günes, "Data clustering," *Intelligent multidimensional data clustering and analysis*, vol. 31, no. 3, pp. 28–72, 2016, doi: 10.4018/978-1-5225-1776-4.ch002.

[31] X. Wang, K. Smith, and R. Hyndman, "Characteristic-based clustering for time series data," *Data Min Knowl Discov*, vol. 13, no. 3, pp. 335–364, Nov. 2006, doi: 10.1007/s10618-005-0039-x.

[32] R. E. Caraka, R. C. Chen, S. W. Huang, S. Y. Chiou, P. U. Gio, and B. Pardamean, "Big data ordination towards intensive care event count cases using fast computing GLLVMS," *BMC Med Res Methodol*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12874-022-01538-4.

[33] R. Silipo, I. Adae, A. Hart, and M. Berthold, "Seven Techniques for Dimensionality Reduction," *KNIME.com*, 2014.

[34] A. Buchholz, A. Buchholz, and S. T. Université, "High dimensional Bayesian computation," ENSAE-CREST Paris, 2018.

[35] Z. Shun and P. McCullagh, "Laplace Approximation of High Dimensional Integrals," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1995, doi: 10.1111/j.2517-6161.1995.tb02060.x.

[36] A. Gongal, S. Amatya, M. Karkee, Q. Zhang, and K. Lewis, "Sensors and systems for fruit detection and localization: A review," *Computers and Electronics in Agriculture*, vol. 116. Elsevier B.V., pp. 8–19, Aug. 01, 2015. doi: 10.1016/j.compag.2015.05.021.

[37] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief Bioinform*, vol. 19, no. 6, pp. 1236–1246, May 2017, doi: 10.1093/bib/bbx044.

[38] H. Sasana and J. Aminata, "Energy subsidy, energy consumption, economic growth, and carbon dioxide emission: Indonesian case studies," *International Journal of Energy Economics and Policy*, vol. 9, no. 2, pp. 117–122, 2019, doi: 10.32479/ijeep.7479.

[39] K. Löfgren and C. W. R. Webster, "The value of Big Data in government: The case of 'smart cities,'" *Big Data Soc*, vol. 7, no. 1, Jan. 2020, doi: 10.1177/2053951720912775.

[40] L. G. B. Ruiz, M. C. Pegalajar, R. Arcucci, and M. Molina-Solana, "A time-series clustering methodology for knowledge extraction in energy consumption data," *Expert Syst Appl*, vol. 160, Dec. 2020, doi: 10.1016/j.eswa.2020.113731.

[41] A. W. Moore, *K-means and Hierarchical Clustering. In Statistical Data Mining Tutorials* . 2001.

[42] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding. Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms.," 2007.

[43] D. Delforge, A. Watlet, O. Kaufmann, M. van Camp, and M. Vanclooster, "Time-series clustering approaches for subsurface zonation and hydrofacies detection using a real time-lapse electrical resistivity dataset," *J Appl Geophy*, 2021.

[44] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," in *Physics Procedia*, Elsevier B.V., 2016, pp. 507–512. doi: 10.1016/j.procs.2016.02.095.

[45] S.-S. Yu, S.-W. Chu, C.-M. Wang, Y.-K. Chan, and T.-C. Chang, "Two improved k-means algorithms," *Appl Soft Comput*, vol. 68, pp. 747–755, Jul. 2018, doi: 10.1016/j.asoc.2017.08.032.

[46] X. Huang, Y. Ye, L. Xiong, R. Y. K. Lau, N. Jiang, and S. Wang, "Time series k-means: A new k-means type smooth subspace clustering for time series data," *Inf Sci (N Y)*, vol. 367–368, pp. 1–13, Nov. 2016, doi: 10.1016/j.ins.2016.05.040.

[47] W. S. Manjoro, M. Dhakar, and B. K. Chaurasia, "Operational analysis of k-medoids and k-means algorithms on noisy data," *International Conference on Communication and Signal Processing (ICCSP)*, pp. 1500–1505, 2016.

[48] A. Saxena *et al.*, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017, doi: 10.1016/j.neucom.2017.06.053.

[49] M. Charrad, N. Ghazzali, U. Laval, and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," 2014. [Online]. Available: http://www.jstatsoft.org/

[50] S. Saref and S. Duygu, "Big Data: A Review," *International Conference on Collaboration Technologies and Systems*, 2013.

[51] A. Starczewski and A. Krzyżak, "Performance Evaluation of the Silhouette Index," 2015, pp. 49–58. doi: 10.1007/978-3-319-19369-4_5.