Journal of Physics: Theories and Applications E-ISSN: 2549-7324 / P-ISSN: 2549-7316
J. Phys.: Theor. Appl. Vol. 5 No. 2 (2021) 92-103 doi: 10.20961/jphystheor-appl.v5i2.59133

Analysis of Earthquake Activity in Indonesia by Clustering Method

Adi Jufriansah[1], Yudhiakto Pramudya[2], Azmi Khusnani[3], Sabarudin Saputra[4]

[1,3]Department of Physics Education, IKIP Muhammadiyah Maumere, Jl. Jend. Sudirman, Waioti, Maumere, NTT

[2]Department of Education Master of Physics, Universitas Ahmad Dahlan, Jl. Pramuka, Umbulharjo, Yogyakarta, DIY

[4]Department of Master of Informatics Engineering, Universitas Ahmad Dahlan, Jl. Pramuka, Umbulharjo, Yogyakarta, DIY

Email : saompu@gmail.com

Received 4 July 2021, Revised 1 August 2021, Published 30 September 2021

Abstract: Indonesia is an area where three large tectonic plates meet, namely the Indo-Australian, Eurasian and Pacific plates, so that Indonesia is included in the earthquake-prone category, with 11,660 earthquake vibrations identified in the Meteorology, Climatology and Geophysics Agency (BMKG) database in 2019 The purpose of this study is to develop a classification of the distribution of earthquakes in Indonesia in 2019 based on the values of magnitude, depth, and position. This research was conducted by using the clustering method based on the K-means algorithm and the DBSCAN algorithm as a comparison. The results of the clustering show that the earthquake data analysis using the K-Means algorithm is superior with a silhouette index value of 0.837, while the DBSCAN algorithm has a silhouette index value of 0.730.

Keywords: Earthquake, Clustering Method, K-means, DBSCAN, BMKG

1. Introduction

Indonesia is located above the connection of the Pacific, Eurasian, and Indo-Australian tectonic plates which continue to move actively so that it is prone to earthquakes due to the release of seismic waves on rocks in the earth's crust (Halim & Widodo, 2017; Kurmiati et al., 2021; Sari et al., 2012). Another trigger for earthquakes in Indonesia is the volcanic activity of active volcanoes surrounding the Indonesian archipelago (Murdiaty et al., 2020). The Meteorology, Climatology and Geophysics Agency recorded earthquakes reaching 400 times every month until in 2019 11,660 earthquakes were recorded on the Earthquake Repo site (Kurmiati et al., 2021).

An earthquake is a shaking event on the earth's surface caused by a sudden release of energy to create seismic waves that hit rocks in the earth's crust (Bahri & May, 2019). Earthquake events are recorded based on location in the form of latitude and longitude and their depth with a certain level of earthquake strength (magnitude) (Akbar et al.,

2018). Until 2019 technological developments have not been able to predict exactly where and when an earthquake will occur even though the location prone to occurrence and the impacts caused by the earthquake have been mapped based on the level of strength recorded through seismographs (Bahri & May, 2019).

Data on the quantity of earthquakes that occur very much every year can be used as the basis for processing earthquake distribution with the data mining method which is stated as a large-scale data processing method to get new information that is easy to understand (Reviantika et al., 2020). Data such as the location or point of the earthquake, the depth level of the epicenter, and the strength of the earthquake can be used as data mining objects for analysis purposes in many relevant studies (Ismail, 2021; Reviantika et al., 2020).

Various analyzes of earthquakes have been carried out using various methods to determine the distribution of earthquakes, earthquake-prone areas, and the impact of earthquakes based on these data. Earthquake analysis can be done using an area classification approach, Ismail (2021) states that the classification of earthquake areas can be done using a random forest algorithm based on earthquake events in the form of coordinates (latitude, longitude), depth (depth), and magnitude (seismic energy strength of the earthquake), with an accuracy of 99.97%. Other methods that can be used in earthquake analysis are K-Means Clustering (Reviantika et al., 2020; Murdiaty et al., 2020) and Business Intelligence methods (Akbar et al., 2018).

K-Means is a type of algorithm in the data mining clusterization method (Reviantika et al., 2020). Analysis using K-Means can provide results in the form of classification data for grouping the distribution of earthquakes, so in this study an analysis of the distribution of earthquakes in Indonesia in 2019 will be carried out based on data on earthquake point locations, earthquake depth levels, and earthquake strength. In addition, this study attempts to compare the clustering method using the K-Means algorithm with DBSCAN in determining the silhouette index. The silhouette index value given shows a statistical measure to choose the optimal number of clusters that can display graphics regarding the accuracy of the placement of an object in a cluster (Nicolaus et al., 2016).

2. Experimental

This study uses real time earthquake data in Indonesia in 2019 obtained from the Meteorology, Climatology and Geophysics Agency (BMKG) database. The data used in this study consisted of latitude, longitude, earthquake magnitude and depth of earthquake data. The analysis in this study uses the K-Means and DBSCAN algorithms. The stages of this research include preparing data in .csv form, then the data is prepared to avoid data that does not provide information by checking for missing data. The next step is to determine the number of clusters, determine the centroid and visualize it (figure 1).

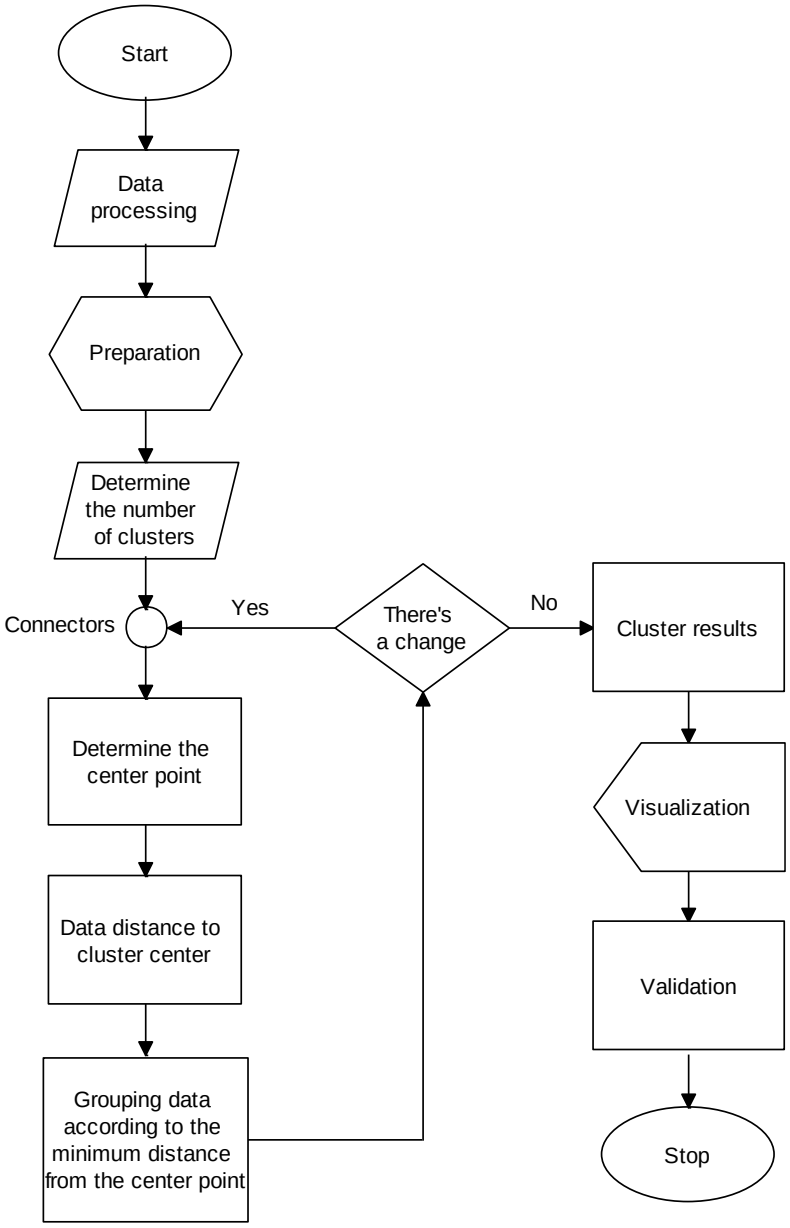


Figure 1. Stages of analysis of earthquake distribution

3. Results and Discussion

Clustering is an unsupervised learning method by grouping data based on the level of similarity without supervision or clustering partition category method (Humairah & Rasyidah, 2020). This method is used because it is more efficient, such as removing redundant variables using correlation and ignoring target variables. The basic principle of clustering is to maximize the similarity between members of one cluster and minimize the similarity between members of different clusters. Clustering can also group data based on the level of similarity and level of accuracy (Kurniati et al., 2021; Syakur et al., 2018). The distance of the data is determined using the equation,

$$d_{ij} = \sqrt{\sum_{k=0}^{n} \left(x_{ik} - x_{jk} \right)^2} \tag{1}$$

The formula for calculating the distance between two points in one dimension, two dimensions, and three dimensions respectively is shown in equation (2) to equation (4) (Siregar, 2018),

$$d_{ij} = \sqrt{\left(x_i - x_j \right)^2 + \left(y_i - y_j \right)^2} \tag{2}$$

$$d\left(p,q \right) = \sqrt{\left(p_1 - q_1 \right)^2 + \left(p_2 - q_2 \right)^2} \tag{3}$$

$$d\left(p,q \right) = \sqrt{\left(p_1 - q_1 \right)^2 + \left(p_2 - q_2 \right)^2 + \left(p_3 - q_3 \right)^2} \tag{4}$$

In this study, data was obtained from the BMKG database in 2019 with a total of 11660 data records of earthquake vibrations. The earthquake distribution database for the first five (5) data is shown in table 1.

Table 1. The first five data

Date	Time	Latitude		longitude		Depth	Magnitude	Moment tensor	Region
12/31/2019	23:03:33.474	5.77	S	104.99	E	83	3.3	-	Southern Sumatra, Indonesia (Status: confirmed)
12/31/2019	22:13:21.681	1.62	N	126.37	E	10	3.9	-	Northern Molucca Sea (Status: confirmed)
12/31/2019	21:33:50.385	3.42	S	99.62	E	35	4.0	-	Southwest of Sumatra, Indonesia (Status: confirmed)
12/31/2019	21:25:05.256	10.15	S	115.94	E	10	3.7	-	Siuth of Bali, Indonesia (Status: confirmed)
12/31/2019	18:23:13.053	8.09	S	107.60	E	18	2.8	-	Java, Indonesia (Status: confirmed)

From table 1, the feature selection is then carried out using only the latitude, longitude, earthquake magnitude and depth of earthquake data attributes as mandatory data to be analyzed. The data is then cleaned using imputation to avoid missing data that can affect machine learning work, as shown in Figure 2. Imputation is used in estimating a data distribution parameter and remains dominantly used in testing new tests (Dempster & Rubin, 1997). This method is an alternative to least squares by maximizing the likelihood function (likelihood) or (log-likelihood). The probability function of the linear model is,

$$L\left(\beta_0, \beta_1, \sigma^2 \right) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma} \right)^2} \right) \tag{5}$$

The maximum likelihood method estimates the parameters β and σ by obtaining the parameter values β_0, β_1 and σ that maximize L (Dempster & Rubin, 1997).

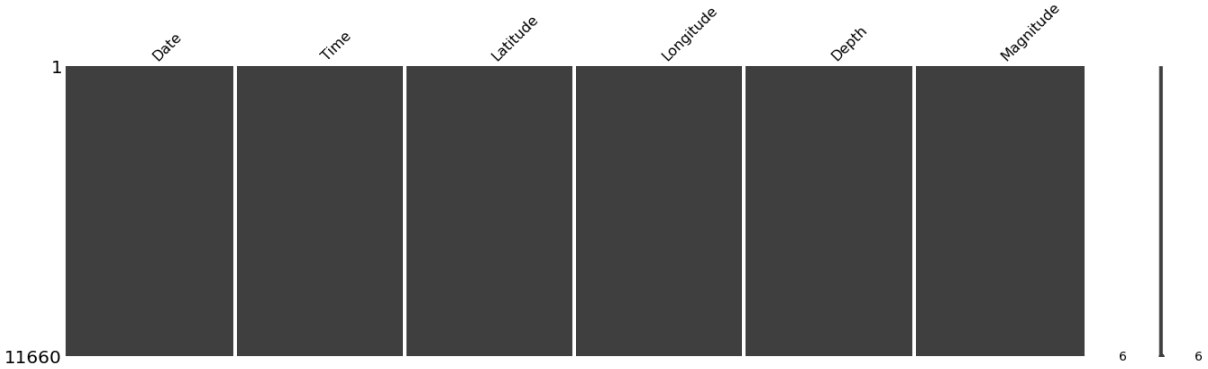


Figure 2. Data cleaning results

Based on Figure 2, it can be explained that the results of feature selection using data cleaning have been successfully carried out. This is indicated by a dominant black color block in each of its attributes. The next step is to calculate the correlation between attributes as shown in Figure 3. The correlation calculation aims to determine the relationship between the variables.

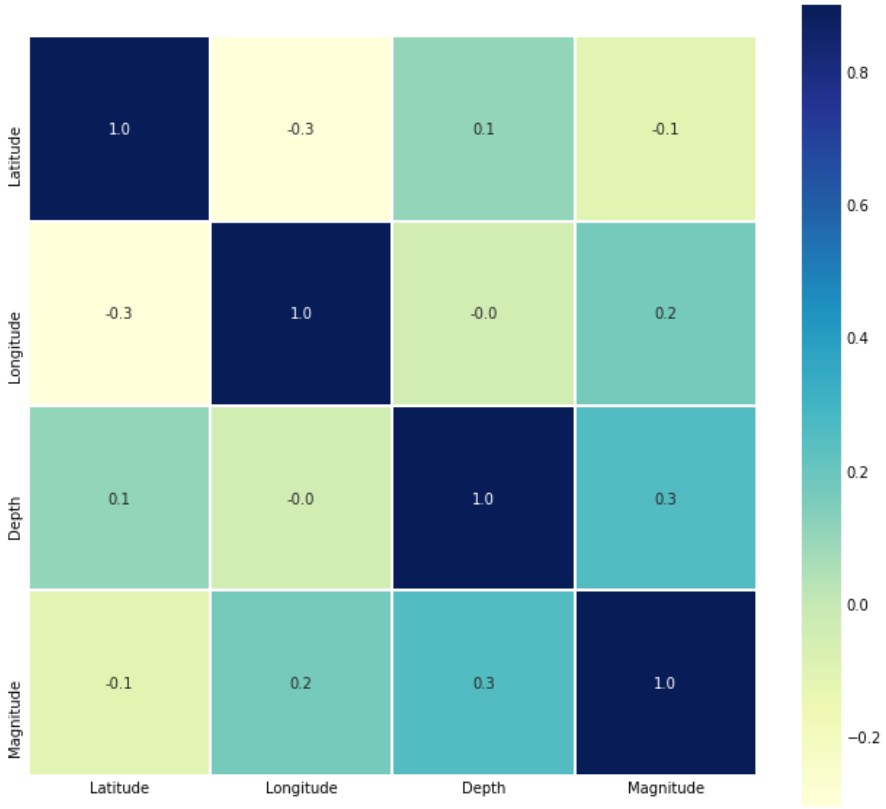
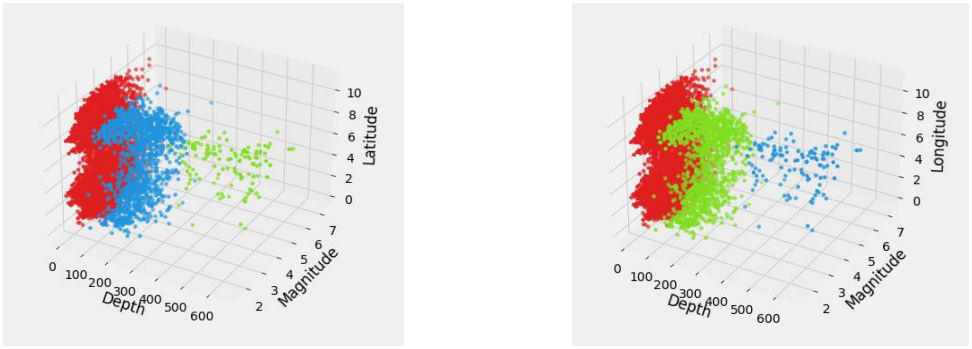


Figure 3. Correlation results with heatmap

The highest correlation result in Figure 3 is 0.3. This shows that the correlation criteria are still weak, so it is necessary to normalize the data. Normalization aims to eliminate or reduce data so as to produce data that matches the expected value. Normalization of data is done using equation (6).

$$x' = \frac{x - \mu}{\sigma} \tag{6}$$

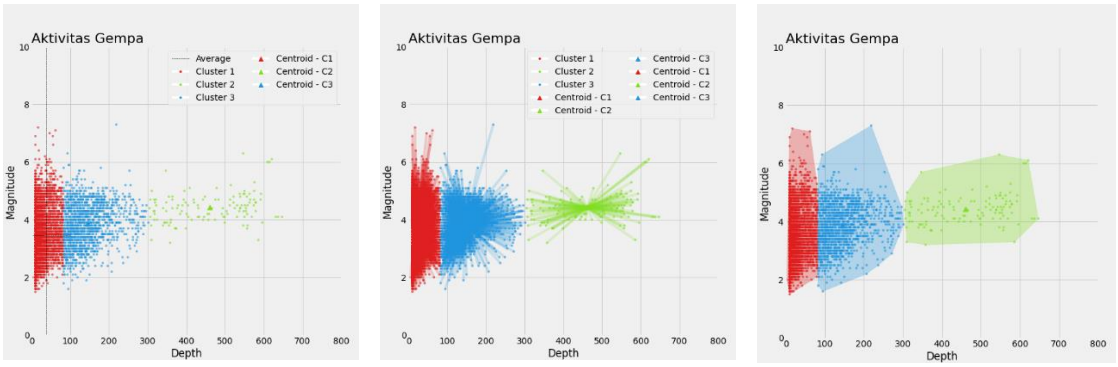
3-D visualization of latitude, longitude, earthquake magnitude and depth data is presented in Figure 4. Figure 4 shows that the distribution of earthquake data is divided into three groups, with Figure 4a for latitude and Figure 4b for longitude..



(a) (b)

Figure 4. Result of 3-D plot of earthquake activity, (a). The results of the latitude, depth and magnitude plots, (b). Results of plotting longitude, depth and magnitude

The results of the earthquake distribution that have been identified, then determined the initial value of the centroid at random. This is useful for calculating the distance of the distribution matrix, so that it can be continued for the stage of grouping objects and determining cluster members according to the minimum distance from the centroid. In addition to this, repeated iterations of the data are carried out in order to produce a new, better centroid distance, as shown in Figure 5.



(a) (b) (c)

Figure 5. (a) Centroid area, (b) Minimum distance to centroid point, and (c) Cluster area

By doing a comparison between Figure 4 and Figure 5, further information is obtained that earthquake activity in Indonesia in 2019 is more common in cluster 1, namely at a depth of 0 km to 90 km and cluster 2 at a depth of 90 km to 300 km. Meanwhile, in cluster 3 for a depth of 300 km to 700 km, fewer occurrences were recorded. This is in accordance with the results of the conference held by ITB in 2021 which is presented in Figure 6.

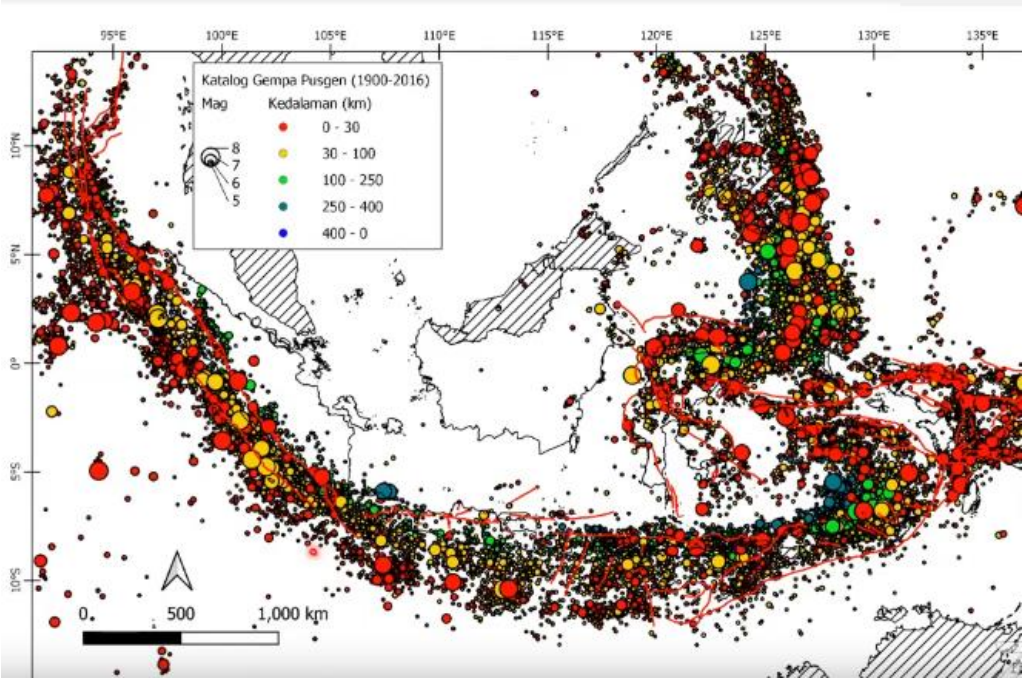


Figure 6. Earthquake catalog (1900-2016) (Permana, 2021)

Figure 6 is the distribution of earthquakes in areas in Indonesia in the last 20 years (1900-2016). In general, the distribution of earthquakes occurred in almost all parts of Indonesia. In general, the generator of earthquakes is the presence of faults and subduction collisions of the earth's zone. Faults cause discontinuity in the rock so that there is a shift. The larger the rock shift area, the greater the resulting magnitude. The distribution of earthquake data in 2019 belongs to a phase that has its own characteristics, this is because before 2019 a large earthquake has been confirmed with an average of ± 7.3 Mw, including the Aceh earthquake in 2004 with 9.2 Mw (Meltzner et al., 2006) , 2006 Yogyakarta earthquake with 6.2 Mw (Sarah & Soebowo, 2013), Lombok earthquake series in 2016 with 6.2 Mw, 2017 earthquake with a scale of II-III, 2018 with 7.0 Mw (Kencanawati et al., 2020), Palu Earthquake 2018 with 7.5 Mw (Mason et al., 2021). So that this event is used as a factor to determine the distribution of the 2019 earthquake.

Based on the results of the clusterization of the earthquake distribution, the centroid values are obtained which are presented in table 2.

Table 2. The centroid of each earthquake attribute

Cluster	Latitude	Longitude	Depth	Magnitude
1	4.21313461	121.40729405	18.64426488	3.40164564
2	4.73856731	121.82910727	146.01655868	3.85773938
3	5.77926829	123.2396748	461.05691057	4.42439024

If presented on a map, the average centroid is on the Eurasian plate in Figure 7 with a blue circle marked.

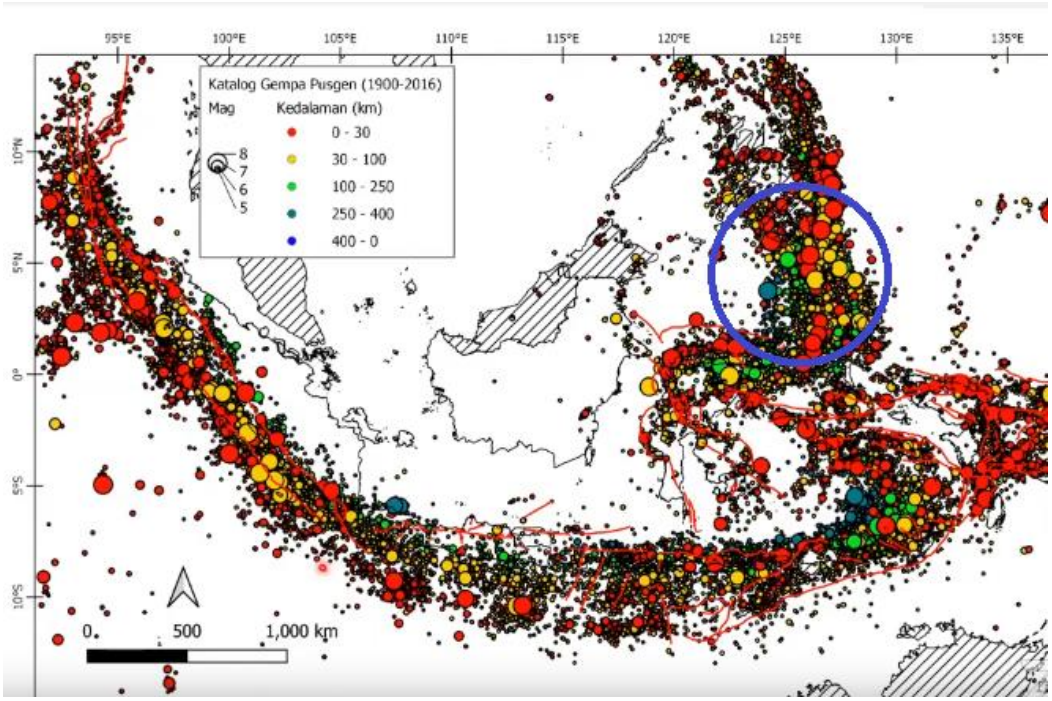


Figure 7. Earthquake centroid position based on clusterization results

The percentage variance of earthquake distribution is explained as a function of the number of clusters. The first cluster will provide a lot of information about the effect of the angle so that it forms an angle, see figure 8. This is in accordance with the data plot generated by each attribute.

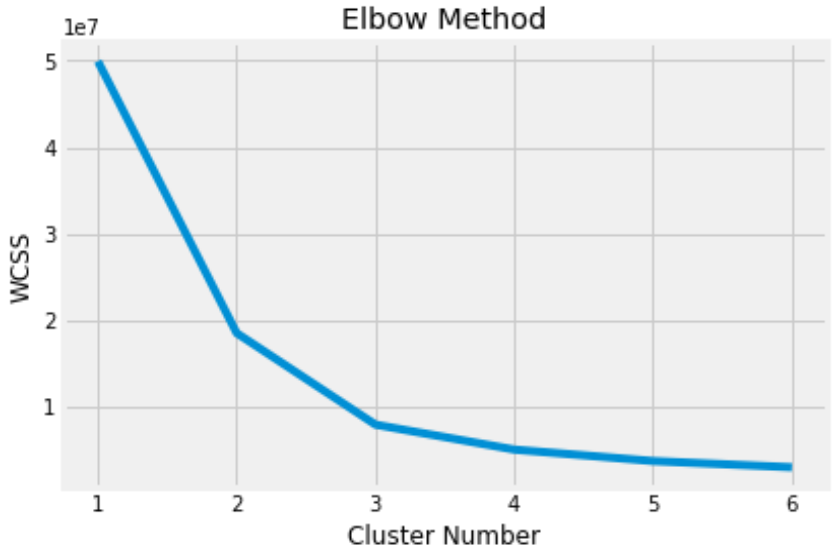


Figure 8. The results of the analysis using the Elbow method

The number of clusters obtained based on data fractures using the Elbow method in Figure 8 is k = 3 or there are 3 clusters, this number is the result of optimal cluster formation for earthquake distribution data in 2019 (Bhoowalia & Kumar, 2014 and Marutho et al., 2018). So that the cluster data output for the first 10 data is shown in table 3.

Table 3. The first ten data using the Elbow Method

Latitude	Longitude	Depth	Magnitude	cluster
5.77	104.99	83	3.3	2
1.62	126.37	10	3.9	1
3.42	99.62	35	4.0	1
10.15	115.94	10	3.7	1
8.09	107.60	18	2.8	1
9.06	114.47	55	3.2	1
8.17	114.89	10	3.3	1
3.38	128.38	12	2.2	1
8.14	107.94	32	2.5	1
2.90	130.29	18	3.5	1
0.01	123.45	104	4.4	2
4.47	124.99	318	4.2	3

Silhouette coefficient value is a statistical measure to choose the optimal number of clusters (Nicolaus et al., 2016). Silhouette value determination can be done using the K-Means algorithm and the DBSCAN algorithm as shown in Figure 9 and Figure 10.

```
# Menjalankan K-Means Clustering ke dataset
kmeans = KMeans(n_clusters = 3, init = 'k-means++', random_state = 42)
y_kmeans = kmeans.fit_predict(x)

print("Silhouette Coefficient: %0.3f"
      % metrics.silhouette_score(x, y_kmeans))
```

Figure 9. K-Means Algorithm

```
# Number of clusters in labels, ignoring noise if present.
n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
n_noise_ = list(labels).count(-1)

print("Silhouette Coefficient: %0.3f"
      % metrics.silhouette_score(x, labels))
```

Figure 10. DBSCAN Algorithm

Based on the results of the analysis, the silhouette value is 0.867 for the K-Means algorithm, while the DBSCAN algorithm is 0.730. So that it can be obtained information that the statistical analysis for the number of clusters with the K-Means algorithm is higher than the DBSCAN algorithm, but both have shown large values. Therefore, the number of earthquake distribution clusters for k=3 is the most optimal. Overall, data testing using clustering and using the Elbow Method on earthquake data in Indonesia in 2019 was appropriate. The K-Means Clustering process uses the Elbow method to determine the best k optimization value. The results of the clusters formed will be labeled

to facilitate the division of each cluster area by considering the characteristics of each attribute. Performance testing using 11660 earthquake repo data. The results of the Sum of Square Error calculation for each cluster experienced the largest decrease at k = 3, which can be seen in Figure 8. This test will look for the performance of each cluster number which is adjusted to the range of values in the Elbow Method. In Figures 5 and 7, information is obtained that the strength of the earthquake magnitude is spread over the depth of each cluster. This research is also supported by the silhouette index of the K-Means algorithm which is compared with the DBSCAN algorithm.

4. Conclusion

The distribution of earthquakes in Indonesia in 2019 recorded 11,660 vibrations obtained from BMKG data from January 1, 2019 to December 31, 2019. The distribution of earthquakes was analyzed based on 4 attributes, namely latitude, longitude, depth, and magnitude data. The results of the analysis obtained that the level of correlation between attributes is still weak so it is necessary to normalize the data. This study also presents a visualization of the earthquake distribution in 3-D form and the results of the centroid area with a value of k = 3 using the K-means algorithm. The cluster area is divided into colors that have been determined by the minimum distance from the distribution of the earthquake point to the centroid. From the average centroid of each attribute, it can be grouped that the distribution point of the earthquake in 2019 is on the Eurasian plate. Clustering data with a value of k = 3 is also strengthened using the Elbow Method which shows the appropriate optimization value for the trial value. The clustering results also show that earthquake data analysis using the K-Means algorithm is superior with a silhouette index value of 0.837, compared to using the DBSCAN algorithm which has a silhouette index value of 0.730.

References

Akbar, R., Darman, R., Marizka, Namora, J., & Ardewati, N. (2018). Implementasi Business Inteligence Menentukan Daerah Rawan Gempa Bumi di Indonesia dengan Fitur Geolokasi. *Jurnal Edukasi Dan Penelitian Informatika*, 4(1), 30–35. https://doi.org/10.26740/jieet.v2n1.p13-18

Bahri, Z., & Mungkin, M. (2019). Penggunaan SCR sebagai Alarm Peringatan Dini pada saat terjadi Gempa Bumi. *JET* (*Journal of Electrical Technology*), 4(3), 101–105.

Bholowalia, P., & Kumar, A. (2014). EBK-means: A Clustering Technique Based on Elbow Method And K-Means in WSN. *International Journal of Computer Applications*, 105(9).

Dempster, A.P., Laird, N. M. & Rubin, D.B. (1997). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*(39), 1- 38.

Halim, N. N., & Widodo, E. (2017). Clustering Dampak Gempa Bumi di Indonesia Menggunakan Kohonen Self Organizing Maps. *Prosiding SI MaNIS* (*Seminar Nasional Integrasi Matematika Dan Nilai Islami*), 1(1), 188–194.

Humaira, H., & Rasyidah, R. (2020). Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. *WMA-2*, January.

Ismail. (2021). Klasifikasi Area Gempa Bumi Menggunakan Algoritma Random Forest. *Jurnal Ilmiah Informatika Komputer*, 26(1), 56–64.

Kencanawati, N. N., Agustawijaya, D. S., & Taruna, R. M. (2020). An Investigation of Building Seismic Design Parameters in Mataram City Using Lombok Earthquake 2018 Ground Motion. *Journal of Engineering & Technological Sciences*, 52(5).

Kurmiati, D., Fauzi, M. Z., Ripangi, Falegas, A., & Indria. (2021). Clustering of Earthquake Prone Areas in Indonesia Using K-Medoids Algorithm. *Malcolm: Indonesian Journal of Machine Learning and Computer Science*, 1(1), 47–57.

Kurniati, D., Fauzi, M. Z., Ripangi, R., Falegas, A., & Indria, I. (2021). *Cluster*isasi Daerah Rawan Gempa Bumi di Indonesia Menggunakan Algoritma K-Medoids: Clustering of Earthquake Prone Areas in Indonesia Using K-Medoids Algorithm. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 1(1), 47-57.

Marutho, D., Handaka, S. H., & Wijaya, E. (2018, September). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. *In 2018 International Seminar on Application for Technology of Information and Communication* (pp. 533-538). IEEE.

Mason, H. B., Montgomery, J., Gallant, A. P., Hutabarat, D., Reed, A. N., Wartman, J., ... & Yasin, W. (2021). East Palu Valley flowslides induced by the 2018 MW 7.5 Palu-Donggala earthquake. *Geomorphology*, 373, 107482.

Meltzner, A. J., Sieh, K., Abrams, M., Agnew, D. C., Hudnut, K. W., Avouac, J. P., & Natawidjaja, D. H. (2006). Uplift and subsidence associated with the great Aceh-Andaman earthquake of 2004. Journal of Geophysical Research: Solid Earth, 111(B2).

Murdiaty, M., Angela, A., & Sylvia, C. (2020). Pengelompokkan Data Bencana Alam Berdasarkan Wilayah, Waktu, Jumlah Korban dan Kerusakan Fasilitas Dengan Algoritma K-Means. *Jurnal Media Informatika Budidarma*, 4(3), 744.

Nicolaus, Sulistianingsih, E., & Perdana, H. (2016). Penentuan jumlah cluster optimal pada median linkage dengan indeks validitas silhouette. *Buletin Ilmiah Math. Stat. Dan Terapannya* (*Bimaster*), 5(2), 97–102.

Permana, A. 2021. *Mengenal Gempa Bumi, Sumber, dan Bahayanya*. https://www.itb.ac.id/news/read/57739/home/mengenal-gempa-bumi-sumber-dan-bahayanya

Reviantika, F., Harahap, C. N., & Azhar, Y. (2020). Analisis Gempa Bumi Pada Pulau Jawa Menggunakan Clustering Algoritma K-Means. *Jurnal Dinamika Informatika*, 9(1), 51–60.

Sarah, D., & Soebowo, E. (2013). Liquefaction Due to the 2006 Yogyakarta Earthquake: Field Occurrence and Geotechnical Analysis. *Procedia Earth and Planetary Science*, 6, 383-389.

Sari, A. W., Jasruddin, & Ihsan, N. (2012). Analisis Rekahan Gempa Bumi dan Gempa Bumi Susulan dengan Menggunakan Metode Omori. *Sains Dan Pendidikan*

Fisika, 8(3), 263–268.

Siregar, A. M. (2018). Penerapan Algoritma K-Means untuk Pengelompokan Daerah Rawan Bencana di Indonesia. *INTERNAL* (*Information System Journal*), 1(2), 1-10.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *In IOP Conference Series: Materials Science and Engineering* (Vol. 336, No. 1, p. 012017). IOP Publishing.