Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 13(2) 2020: 194-205
http://ejournal.undip.ac.id/index.php/media_statistika

CLUSTERING OF EARTHQUAKE RISK IN INDONESIA USING K-MEDOIDS AND K-MEANS ALGORITHMS

Isna Hidayatur Rifa, Hasih Pratiwi, Respatiwulan
Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret

e-mail: hpratiwi@mipa.uns.ac.id

DOI: 10.14710/medstat.13.1.194-205

Article Info:
Received: 27 November 2020
Accepted: 19 October 2020
Available Online: 28 December 2020

Keywords:
Earthquake, Data Mining, Clustering, K-Medoids Algorithm, K-Means Algorithm
.

Abstract: Earthquake is the shaking of the earth's surface due to the shift in the earth's plates. This disaster often happens in Indonesia due to the location of the country on the three largest plates in the world and nine small others which meet at an area to form a complex plate arrangement. An earthquake has several impacts which depend on the magnitude and depth. This research was, therefore, conducted to classify earthquake data in Indonesia based on the magnitudes and depths using one of the data mining techniques which is known as clustering through the application of *k*-medoids and *k*-means algorithms. However, *k*-medoids group data into clusters with medoid as the centroid and it involves using clustering large application (CLARA) algorithm while *k*-means divide data into *k* clusters where each object belongs to the cluster with the closest average. The results showed the best clustering for earthquake data in Indonesia based on magnitude and depth is the CLARA algorithm and five clusters were found to have total members of 2231, 1359, 914, 2392, and 199 objects for cluster 1 to cluster 5 respectively.

1. INTRODUCTION

According to the United Nations International Strategy for Disaster Reduction (UNISDR), Indonesia is the country which is most prone to natural disasters in the world (Briceno, 2007). The country occupies a very active tectonic zone which is a complex meeting point between the world's three major plates and nine small others (Bird, 2003). The meeting and movement of these three plates have created a range of active volcanoes and potential earthquakes throughout the Indonesian archipelago (Pratiwi et al., 2018). Moreover, the interactions between these plates also increase the proneness of the country to earthquakes (Milsom et al., 1992).

Several efforts have been made to mitigate earthquake disaster and the first step include grouping the areas of occurrence to determine its potential and characteristics in each region. This is in line with the findings of Febriani and Hakim (2015) that the daily occurrence of earthquakes in Indonesian territory needs a group analysis to determine its

central area and depth. Several studies have, however, been conducted such as Saracli et al. (2013) to discuss the hierarchical group analysis and Arbelaitz et al. (2013) which focused on the validity of grouping results.

The current technological advancement has caused quite rapid developments and this has also improved the data in the fields of science, business, and government. Moreover, internet development has contributed significantly to data accumulation to produce big data but the analysis methodology is unable to handle a large amount of data. Business people and researchers, however, need to take advantage of existing big data, and a new technology known as data mining was developed for this purpose. This concept is the process of extracting or mining knowledge needed by big data and one of the techniques usually used is clustering.

Clustering is a group or cluster formation method which involves placing an object in the same cluster with another related object but occupies a different cluster when it involves an unrelated object. This means the technique focuses on grouping objects based on certain similarities and it is further divided into two which are hierarchical and partitioned clustering. According to Han et al. (2012), the hierarchical aspect classifies data by creating a hierarchy in the form of a dendrogram with similar data placed in adjacent hierarchies while dissimilar ones are in distant hierarchies. Meanwhile, partitioned aspect divides data into *k* parts with each part representing a cluster and some of the methods often used are *k*-means and *k*-medoids.

According to Septiana and Djohan (2015), *k*-means is a method used in dividing the number of observations into *k* clusters and the set of each observation belongs to the cluster with the closest average. It is important to note that an average is defined as the data center measure which has a value affected by extreme values. Meanwhile, *k*-medoids group objects into several clusters using the medoid as the centroid and this makes it more robust than the average centroid.

This background description led to the clustering of the earthquake data in Indonesia using the *k*-medoids and *k*-means algorithms with the depth and magnitude of the earthquake used as the attributes after which the clusters produced were compared to determine the correct partition.

2. LITERATURE REVIEW

2.1. Earthquake

Earthquake is the shaking of the earth's surface by the sudden movement or shift of the rock layers on the earth's crust due to the movement of tectonic plates (Sunarjo et al., 2010). These plates are the elastic rock layer and this means the energy received from the mantle layer is stored in the form of elastic energy. However, in a case the energy received exceeds the elasticity limit, the energy is expected to be released in the form of elastic waves and this causes vibrations or shocks in the earth's layer. Earthquakes are further defined as random and irregular natural phenomena in time and space (Pratiwi et al., 2018) happening without any previous signs.

Fowler (2005) classifies earthquakes based on the depth as follows:

1. shallow earthquakes: less than 70 km,
2. intermediate earthquakes: between 70km and 300 km, and
3. deep earthquakes: more than 300 km.

According to Sunarjo et al. (2010), shallow earthquakes cause shock and destruction effects which are more devastating than deep ones due to the closeness of the source to the earth's surface, and this further increases its wave energy.

2.2. Data Mining

Data mining is a process of using one or more machine learning techniques to analyze and extract knowledge automatically. It is also defined as induction-based learning which is the process of forming general definitions for a concept by observing specific examples to be studied (Hermawati, 2013) Moreover, data mining is an iterative and interactive process of finding valid or perfect, useful, and understandable new patterns or models in a very large database.

Several analysis techniques have been observed in data mining and one of them is clustering which has been previously reported to be used in dividing a set of data into several subsets or clusters. Consequently, there is, however, a high degree of similarity in the elements of a particular cluster and a low level of similarity between different clusters. Meanwhile, clustering is also called unsupervised learning and this means a data mining technique which does not require initial training of data.

Clustering is a group or cluster formation method which involves placing an object in the same cluster with another related object but occupies a different cluster when it involves an unrelated object. The purpose of this technique is to minimize the distance within the cluster and maximize the distance between clusters. Moreover, hierarchical clustering is a set of clusters arranged as a hierarchical tree while partitioned clustering is the division of data objects into non-overlapping clusters to ensure each data object is in exactly one subset.

2.3. K-Medoids Algorithm

K-medoids is a part of the partitioning method algorithm which has been reported to be more robust compared to the data clusters containing outliers. The algorithms often used in this technique include Partitioning Around Medoids (PAM) and Clustering Large Application (CLARA). The PAM algorithm works effectively for small data but less effective for large data while the sampling-based algorithm, CLARA, is usually applied to large datasets using the following steps stated by Kaufman and Rouesseeuw (1987) :

1. Determining *k* which is the number of clusters to be formed,
2. dividing the dataset into multiple subsets of fixed size and the sample size was observed to be minimal,
3. choosing the center of the cluster or the medoid as much as *k*,
4. calculating the distance between non-medoid objects and the medoid in each cluster and placing each non-medoid object to the nearest medoid after which the total distance is calculated,
5. randomly selecting non-medoid objects in each cluster as candidates for the new medoid,
6. calculating the distance of each non-medoid object to the new candidate and assigning each non-medoid object to the closest candidate after which the total distance is calculated,
7. calculating the difference between the total distance (*S*) which is the difference between the total distance between the old medoid and the new candidate,
8. if a value is obtained, the candidate medoid becomes the new medoid, and
9. repeating steps 5 to 8 until there is no change in the medoid.

2.4. K-Means Algorithm

K-means is a partitioned clustering with each cluster connected by a centroid or center point and each point is placed into a cluster with the closest centroid. The steps used in this algorithm according to Han and Kamber (2012)include

1. determining the number of clusters *k*,
2. selecting the centroid of the objects to be grouped as much as *k*,
3. determining the distance of each object to each centroid by calculating the distance using a similarity measure,
4. allocating each object to the nearest centroid and determining the new centroid by calculating the average of the data in each cluster using the following formula

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

where μ_k is the new centroid in the *k*-th cluster, N_k is the numbers of data in the *k*-th cluster, and x_q is the *q*-th data in the *k*-th cluster, and

5. repeating steps 3 to 4 until there is no change in centroid.

2.5. Silhouette Coefficient Method

The silhouette coefficient method was used for evaluation by testing the quality of the resulting clusters. This method combines two others which are the cohesion and separation methods with the cohesion used to measure the closeness of the data in one cluster while separation evaluates the proximity between the clusters formed (Arbelaitz et al., 2013).

The silhouette of a cluster is a plot of all members' silhouettes sorted in descending order with the highest value for each cluster placed at the top to compare the quality based on the width such that a wider silhouette produces a better cluster quality. The silhouettes for the *k*-medoids were conducted several times using different k values to compare the plots produced and the *k* with the highest average width was called the silhouette coefficient plot (Kaufman & Rousseeuw, 1987). Meanwhile, the steps involved in the method as stated by Struyf et al. (1997) are as follows:

1. Calculate the average distance between the *i*-th object and all objects in cluster A using the following equation

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

where *j* is another object in cluster A and $d(i,j)$ is the distance between objects *i* and *j*.

2. Calculate the average distance between the i-th object and another in the other cluster using the following equation

$$d(i,C) = \frac{1}{|C|} \sum_{j \in C} d(i,j)$$

where $d(i,C)$ is the average distance between *i*-th object and others in other cluster C with $A \neq C$ where A is the number of cluster A members and C is the number of cluster C members.

3. The minimum average object distance value, $b(i)$, which shows the average difference between object i for the cluster closest to its neighbor was determined using the following equation

$$b(i) = \min_{C \neq A} d(i, C)$$

4. Calculate the silhouette value with the equation

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

The results of $s(i)$ are, therefore, expected to be in the range of -1 to 1 and the values can be interpreted as follows in accordance with Kaufman and Rouesseeuw (1987):

a. $s(i) \approx 1$ means the object *i* is located in the proper cluster (*A*),
b. $s(i) \approx 0$ means the object *i* is located between two clusters (*A* and *C*),
c. $s(i) \approx -1$ means the object *i* is located in an improper cluster which is closer to B than A.

5. Calculate the silhouette coefficient (SC) defined as average using the following formula

$$SC = \frac{1}{n} \sum_{i=1}^{n} s(i)$$

where *n* is the number of observations.

3. RESEARCH METHOD

Indonesian earthquake data from 1973 to 2017 obtained from the United States Geological Survey (USGS) which is a scientific agency of the United States government were used in this research with the attributes emphasized being the magnitude and depth of the earthquake.

The steps used in the research include:

1. Determination of the number of clusters *k* to be formed using the silhouette method as described in Section 2.5.
2. Clustering of data using the CLARA algorithm:
   a. The distance of each object to each medoid was calculated using the size of Euclid's distance in each cluster and each object was placed to the nearest medoid after which the total distance was calculated,
   b. a new medoid candidate was selected,
   c. the distance of each object to each medoid was calculated using the size of Euclid's distance in each cluster and each object was placed to the nearest medoid after which the total distance was calculated,
   d. the difference between the total distances was calculated,
   e. the value obtained was used to name the candidate as the new medoid, and
   f. steps c to f were repeated until there was no change in the medoid.
3. The results of the clustering were displayed using the *k*-medoids algorithm.
4. Data clustering was conducted using the *k*-means algorithm
   a. The distance of each data from each centroid was calculated using the appropriate distance measurement based on step 4,
   b. each data was classified based on its proximity to the centroid or smallest distance
   c. the centroid value was updated with the new centroid value obtained from the mean of the cluster concerned, and
   d. steps a to d were iterated until no change was observed in each cluster member.
5. The results of the clustering were displayed using the *k*-means algorithm,
6. The clusters obtained were compared using the silhouette coefficient average value and a greater value has been reported to have better cluster quality.

4. RESULTS AND DISCUSSION

4.1. Data Description

Indonesian earthquake data consists of 7095 earthquake points with two attributes which are depth and magnitude and the depth used was based on the shallow earthquake classification with a maximum of 70km. Some of these data are shown in Table 1 and the depth attribute is the absolute value in the original data obtained from the USGS web reduced by 70. Therefore, the effect produced is the same between the magnitude and depth attributes, and the greater values were observed to be producing a higher level of risk caused by an earthquake.

Table 1. Earthquake Data in Indonesia

Location	Depth (km)	Magnitude (mb)
0 km west of Komerda, Indonesia	42.00	6.47
100 km south of Kotaagung, Indonesia	24.07	5.10
100 km south of Sungaipenuh, Indonesia	14.70	5.84
100 km southeast of Bengkulu, Indonesia	19.52	5.00
100 km southwest of Cigarogol, Indonesia	43.00	6.09
100 km southwest of Sibolga, Indonesia	34.00	6.09

On the USGS website, the name of the place where the earthquake happened is used as the row name and the summary is presented in Table 2.

Table 2. Earthquake Data Summary in Indonesia

	Depth (km)	Magnitude (mb)
Min	0.00	2.500
Quartile 1	30.00	5.100
Median	37.00	5.400
Mean	35.43	5.441
Quartile 3	41.00	5.837
Max	69.20	8.231

The information on the table shows that the maximum and minimum value ranges between the attributes are quite different as observed in the minimum of 0 for depth and 2.5 for magnitude while a maximum of 69,2 and 8,231 respectively were also recorded.

Table 3. Earthquake Standardization Data in Indonesia

Location	Depth (km)	Magnitude(mb)
0 km west of Komerda, Indonesia	0,4668959	1,4992631
100 km south of Kotaagung, Indonesia	-0,8069130	-0,4980922
100 km south of Sungaipenuh, Indonesia	-1,4725900	0,5786882
100 km southeast of Bengkulu, Indonesia	-1,1301607	-0,6442152
100 km southwest of Cigarogol, Indonesia	-0,5379393	0,9469181
100 km southwest of Sibolga, Indonesia	-0,1014516	0,9469181

It is necessary to standardize before clustering to create the same scale for both attributes and the products of the standardization in this research are presented in Table 3 while the summary is indicated in Table 4.

Table 4. Summary of Earthquake Standardization Data in Indonesia

	Depth(km)	Magnitude(mb)
Min	-2.5169	-4.2973
Quartile 1	-0.3856	-0.4981
Median	0.1117	-0.0597
Mean	0.0000	0.0000
Quartile 3	0.3959	0.5787
Max	2.3993	4.0769

The data in Table 4 shows the average data is 0 after standardizing the earthquake data in Indonesia and this means the two attributes have the same scale and the data was applied for the next process.

4.2. Determination of the Number of Clusters

The silhouette method was used in this study to determine the optimal number of clusters and a greater value has been adjudged to be producing the better cluster quality and the plot obtained is presented in Figure 1.

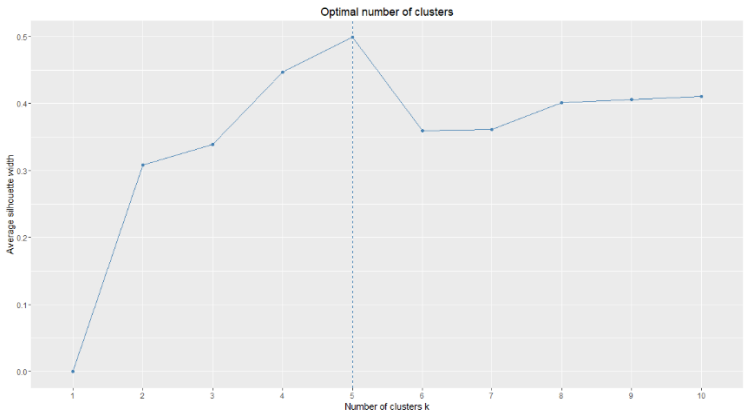

Figure 1. Optimal Number of Clusters Using the Silhouette Method

Figure 1 shows the optimal number of clusters formed is five and the earthquake in Indonesia was further divided into five clusters using the k-medoids and k-means algorithms.

4.3. Clustering Using *K*-Medoids and *K*-Means Algorithms

The *k*-medoids algorithm was used to classify the data into 5 clusters according to the previous step using the CLARA under the R software. The packages were cluster, factoextra, and ggplot2 and the plot produced is shown in Figure 2. The *k*-means algorithm was also used to classify the data based on the distance closest to the average center. The number of clusters used in the *k*-medoids algorithm which involved grouping the data into five clusters ($k = 5$) was also applied and the results of the clustering visualization are indicated in Figure 3.

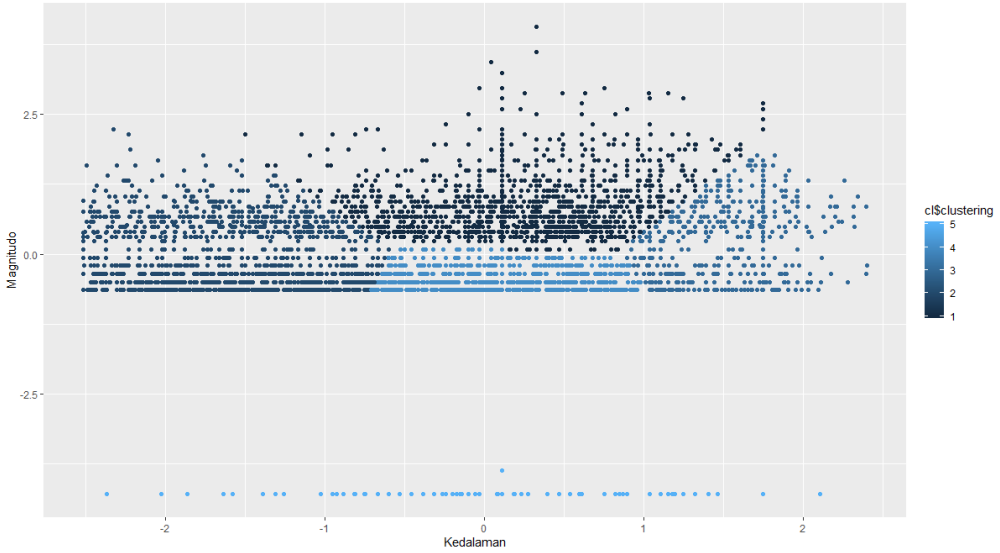

Figure 2. Plot Clustering Using CLARA Algorithm

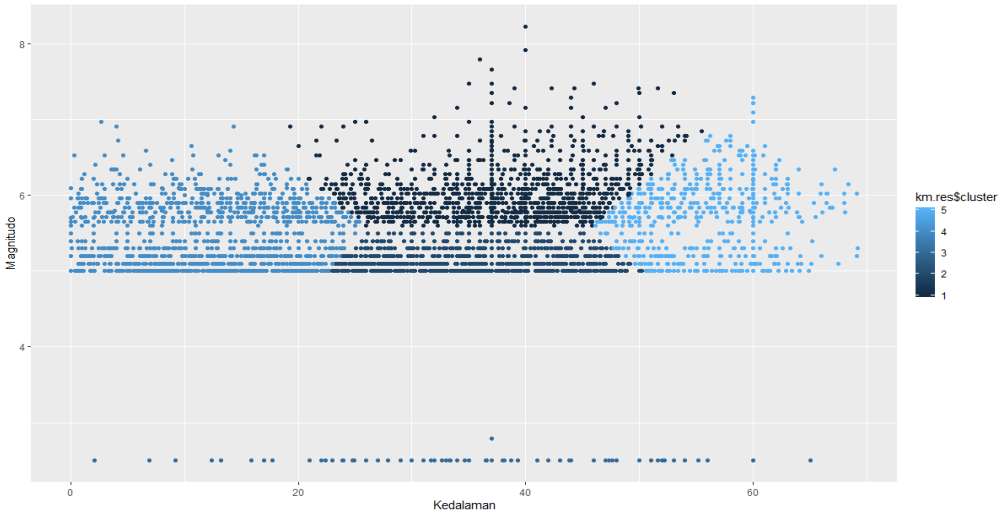

Figure 3. Plot Clustering Using K-Means Algorithm

4.4. Comparison of Clustering Results

The clustering results in Sections 4.3 were compared using the silhouette coefficient method using the average silhouette width from both algorithms and the value for the *k*-medoids using the R software output was found to be 0.546 while *k*-means had 0.516. This means the *k*-medoids algorithm had better results compared to the *k*-means due to its higher average silhouette width value for the data studied. Meanwhile, the results of the cluster center obtained from the CLARA algorithm are shown in Table 5.

Table 5. Cluster Center of CLARA Algorithm

Location	Depth (km)	Magnitude (mb)	*Cluster Members*
Seram, Indonesia 105	37,60	5,90	2231
66 km Southwest of Kotaagung, Indonesia	15,29	5,30	1359
Sulawesi, Indonesia 95	37,00	5,30	914
Talaud Island, Indonesia 67	37,00	5,10	2392
Southwest Sumatra, Indonesia 72	37,00	2,50	199

Table 5 shows the center of the first cluster is in Maneo Rendah Village, North Seram Sub-District, Central Maluku Regency, Maluku Province with the coordinate point of 3°04'44.4"S 129°46'48.0"E, and the second is located in the southwestern Indian Ocean, Kota Agung Sub-District, Tanggamus Regency, Lampung Province with the coordinate point of 6°02'20.0"S 104°21'57.6"E. Moreover, the center of the third cluster is in Tomimi Bay close to Unauna Island, Lembanya SubDistrict, Tojo Una-Una Regency, Central Sulawesi Province with a coordinate point of 0°01'30.0"S 121°38'56.4"E while the fourth is in the Philippine Sea close to the Talaud Islands Regency, North Sulawesi Province with a coordinate point of 3°49'58.8"N 126°30'43.2"E. The center of the fifth cluster is in the Indian Ocean, southwest of Sumatra island with a coordinate point of 6°22'22.8"S 102°19'08.4"E. Furthermore, the first to fifth clusters are shown to have 2231, 1359, 914, 2392, and 199 cluster members respectively.

Cluster 1 has the largest value for the depth and magnitude of the earthquake and this means it has the greatest risk when compared to the other while Cluster 2 has the same magnitude as Cluster 3 but they have different depths. Cluster 3 also has a greater depth in comparison with Cluster 2 and this means it is at greater risk while Cluster 4 was observed to have a moderate risk. Meanwhile, Cluster 5 has the smallest risk level with the smallest magnitude even though it has the same depth as Clusters 3 and 4. Therefore, the risk level of each cluster from the largest to the smallest is 1, 3, 4, 2, and 5.

4.5. Visualization of Clustering Results

The earthquake points in Indonesia according to the risk level order are shown in Figures 4 to Figure 8 with those having high-risk level symbolized by points in Cluster 1, quite high-risk level by points in Clusters 3 and 4, quite low-risk level in Cluster 2 while the points with a low-risk level are indicated by points in Cluster 5.

A similar level was observed around the western and southern coasts of the Sumatra Island, south coast of Java Island, almost all areas on Nusa Tenggara Islands, north and northeastern Sulawesi Island, and the mainland area of Papua Island. Meanwhile, Cluster 5 with a low-risk level has different points from the other clusters which are mostly located around the southwestern coast of Sumatra island and this means the earthquake points in Indonesia are almost the same and completely scattered in every region. The percentage of each cluster based on risk level was respectively 31.44%, 12.88%, 33.71%, 19.15%, and 2.80%. Therefore, Indonesia was found to be dominated by earthquakes with a magnitude $M \geq 5$.

Figure 4. Earthquake Point with High-Risk Level

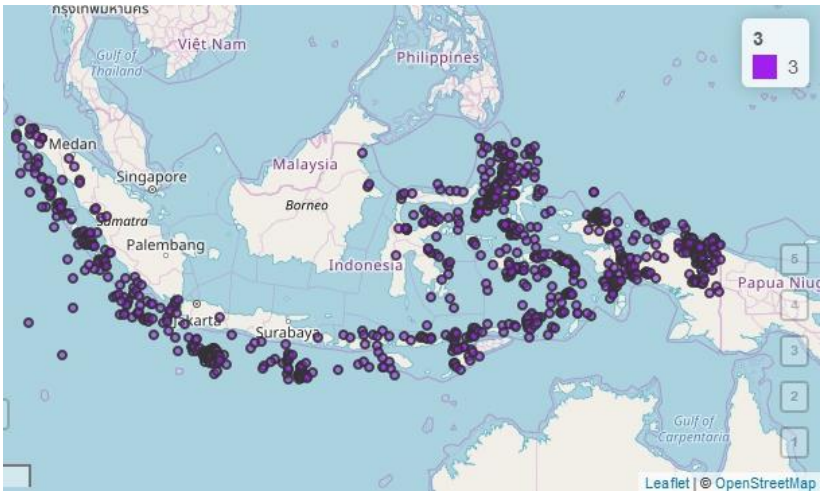

Figure 5. Earthquake Point with a Quite High-Risk Level

Figure 6. Earthquake Point with Moderate-Risk Level

Figure 7. Earthquake Point with a Quite Low-Risk Level

Figure 8. Earthquake Point with Low-Risk Level

5. CONCLUSIONS

The results and discussion showed it is better to use the CLARA algorithm than the *k*-means algorithm in clustering earthquake data in Indonesia in 1973-2017 based on its depth and magnitude. The method was able to produce five clusters with Cluster 1 centered in Maluku Province consisting of 2231 objects observed to have the shallowest depth and the greatest magnitude, thereby, having the greatest risk for earthquake. Meanwhile, Cluster 2 with 1359 objects centered in Lampung Province have a small risk while Cluster 3 with 914 objects centered in Central Sulawesi Province have a high risk and Cluster 4 with 2392 objects centered in North Sulawesi Province was at moderate risk but Cluster 5 consisting of 199 objects centered in the Indian Ocean have the highest depth and lowest magnitude, therefore, it has the lowest risk. This means the level of risk in the clusters in the order 1, 3, 4, 2, and 5 from the highest to the lowest.

The silhouette average value in the *k*-medoids algorithm was found to provide a bond which is strong enough between the object and the formed group. The use of other algorithms, for example the agglomerative hierarchical clustering algorithm or divisive hierarchical clustering, can be done to provide a better result.

REFERENCES

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An Extensive Comparative Study of Cluster Validity Indices. *Pattern Recognition*, *46*(1), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Bird, P. (2003). An Updated Digital Model of Plate Boundaries. *Geochemistry, Geophysics, Geosystems*, *4*(3). https://doi.org/10.1029/2001GC000252

Briceno, S. (2007). *Perkataan Menjadi Tindakan: Panduan untuk Mengimplementasikan Kerangka Kerja Hyogo*. http://www.unisdr.org/ files/594_Bahasa HFA.pdf

Febriani, B. S. & Hakim, R. F. (2015). Analisis Clustering Gempa Bumi Selama Satu Bulan Terakhir dengan Menggunakan Algoritma Self-Organizing Maps (SOMs) Kohonen. *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UMS*, 715–

722.

Fowler, C. M. R. (2005). *The Solid Earth: An Introduction to Global Geophysics* (2nd Edition). Cambridge University Press.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd Edition). Elsevier Inc.

Hermawati, F. . (2013). *Data Mining*. CV Andi Offset.

Kaufman, L. & Rousseeuw, P. J. (1987). *Clustering By Means of Medoids*.

Milsom, J., Masson, D., Nichols, G., Sikumbang, N., Dwiyanto, B., Parson, L., & Kallagher, H. (1992). The Manokwari Trough and the Western End of the New Guinea Trench. *Tectonics*, *11*(1), 145–153. https://doi.org/https://doi.org/10.1029/91TC01257

Pratiwi, H., Rini, L. S., & Mangku, I. W. (2018). Marked Point Process for Modelling Seismic Activity (Case Study in Sumatra and Java). *Journal of Physics: Conference Series*, *1022*(1). https://doi.org/10.1088/1742-6596/1022/1/012004

Saraçli, S., Doğan, N., & Doğan, I. (2013). Comparison of Hierarchical Cluster Analysis Methods by Cophenetic Correlation. *Journal of Inequalities and Applications*, *203*, 1–8. https://doi.org/10.1186/1029-242X-2013-203

Septiana, L., & Djohan, N. (2015). Analisis Perbandingan Algoritma K-Means Clustering dan Expectation-Maximation (EM) untuk Klasifikasi Butir Beras. *Jurnal Teknik Dan Ilmu Komputer*, *4(15)*, 245–253.

Struyf, A., Hubert, M., & Rousseeuw, P. J. (1997). Integrating Robust Clustering Techniques in S-PLUS. *Computational Statistics and Data Analysis*, *26*(1), 17–37. https://doi.org/10.1016/S0167-9473(97)00020-0

Sunarjo, Gunawan, M.T., & Pribadi, S. (2010). *Gempa Bumi Indonesia* (Populer). Badan Meteorologi Klimatologi dan Geofisika. Jakarta.

USGS. *Search Eartquake Catalog*. (n.d.). United States Geological Survey.