Spherical K-Means method to determine earthquake clusters

D S Rini¹, I Sriliana², P Novianti³, S Nugroho⁴, P Jana⁵
¹,²,³,⁴University of Bengkulu, Jl. W.R. Supratman, Bengkulu 38125, Indonesia
⁵Universitas PGRI Yogyakarta, Jl. IKIP PGRI I No. 117, Yogyakarta 55182, Indonesia
dyah.setyorini@unib.ac.id

Abstract. Spherical clustering is a grouping technique for spherical data. A vector data set is grouped into clusters where the distance used to group the vectors is the angle between the vectors. Earthquake events are directional data that can be analyzed in three-dimensional (spherical) form because the epicenter of the earthquake spreads is very long, and the time difference between events is different. Therefore, this study aims to determine earthquake clusters in Bengkulu Province and the surrounding areas. The data used is data on earthquake events in Bengkulu Province and surrounding areas from 1970 to 2019, including latitude, longitude, magnitude, and depth, sourced from the USGS website. The analysis used in determining the earthquake clusters in Bengkulu Province is the Spherical K-means cluster analysis. The results showed that there were six earthquake clusters in Bengkulu Province and its surroundings. Most of the clustering results are in the sea.

1. Introduction

Earthquake is vibrations or shocks that occur on the surface of the earth due to the sudden release of energy that creates seismic waves. Earthquake is caused by the released energy generated by the pressure exerted by a moving plate. When the pressure increases and eventually reaches a state where the pressure can no longer react by the plate, an earthquake will occur. The types of the earthquake are tectonic earthquake, volcanic earthquake, and artificial earthquake.

The tectonic earthquake is the most common type of earthquake and can cause both damage and casualties. This earthquake occurs because of the movement of the earth's tectonic plates that occur suddenly, causing vibrations to the surface of the earth. There are parts of the earth's surface that can be the potential areas for this earthquake. It occurs in the areas where the earth's tectonic plates meet either on land or at sea.

Bengkulu Province is one of the provinces located at the meeting of the Indo-Australian plates and Eurasian plates, which is the main generator of high earthquake activity. The movement of the two plates can cause active faults, which are generators of seismicity in Sumatra. Bengkulu is located between two active faults, namely the Semangko and Mentawai faults. These conditions make Bengkulu Province an earthquake-prone area [1]. Base on the Regional Disaster Management Authority data in Bengkulu Province shows that the epicenter of the earthquake that occurred in Bengkulu was different. The difference between earthquake epicenter and the time of the earthquake occurrence was caused by geology and seismicity in Bengkulu Province. Therefore, each earthquake event has a relationship in time and space.
A measure of the strength of an earthquake based on the energy released from the epicenter is called the magnitude (on the Richter scale). The point in the bowels of the earth, which is the source of the earthquake, is called the hypocenter or focus. Meanwhile, the hypocenter's perpendicular projection to the earth's surface is called the epicenter. The distance between the hypocenter and the epicenter is called the depth of the earthquake. The epicenter states the latitude and longitude positions of the earthquake, so earthquake data are directional data. Directional data is a set of observations in the form of directions in two dimensions (circular) or in three dimensions called spherical data [2].

Earthquake events are directional data that can be analyzed in three-dimensional (spherical) because the epicenter of the earthquake spreads very long, and the time difference between events is different. In addition, the occurrence of earthquakes is influenced by differences in geological conditions and seismicity in the areas at the epicenter. Therefore, an analysis that can classify earthquake events based on the characteristics and spherical proximity of the earthquake points is needed. Statistically, this method is called spherical clustering.

Spherical clustering is a grouping technique for spherical data; a vector data set grouped into clusters where the distance used to group the vectors is the angle between the vectors. Vector data is grouped by direction and not based on the overall length of the vector. The purpose of spherical clustering is to form vector clusters which have the same direction [3].

Weatherill and Burton (2009) conducted research on K-means cluster analysis in shallow seismicity source areas in the Aegean (including Greece, Albania, Former Yugoslav Republic of Macedonia (F.Y.R.O.M.), Southern Bulgaria, and western Turkey). This research develops a K-means partition of seismicity into the source model and estimates the seismotectonic image in the Aegean. In this study, the weighting of the K-means analysis was also carried out with the length of the fault as a weight for each earthquake event data. The results show that the model contains 20 to 30 clusters for the Aegean seismicity source area [4]. In 2016, Hashemi and Karimi model the source of seismicity based on earthquake clusters in the United States. The results of the clustering are used to predict earthquake magnitudes in the United States [5]. Novianti et al. (2017) clustered earthquake events in Bengkulu Province using K-means cluster analysis. The results showed that the optimum number of clusters for earthquake events in Bengkulu Province was 7 clusters which were determined based on the Krzanowski and Lai index [1]. In 2018, Rini et al. examined the optimum number of clusters for earthquake occurrences in Bengkulu Province using internal cluster validation (several internal cluster validation indices), which were 6 clusters based on Davies-Bouldin index [6]. The latest research on directional statistics was conducted by Kim & SenGupta (2018), which introduced Spherical Clustering to group data in three-dimensional directions [3]. As data with spatial elements and direction, earthquake events can be used as spherical data.

2. Spherical K-Means Cluster Analysis

Cluster analysis is a multivariate technique that has the main objective of classifying objects based on their similar characteristics. The characteristics of objects in a cluster have a high level of similarity, while the characteristics between objects in a cluster and other clusters have a low level of similarity. In other words, the diversity within a cluster is minimum, while the diversity between clusters is maximum.

MacQueen suggested using K-means to describe the algorithm in determining an object into a particular cluster based on the nearest distance of the center (means). This process consists of three stages [7]:

1. Partition objects into initial cluster K
2. Starting with recording objects, determine an object into a cluster that its distance close to the center (mean). Distance is usually calculated using Euclid's distance with standardized or non-standardized observations. Recalculate the centroid for the cluster that received the new object and for the cluster
that lost the object. Centroid in one group is calculated by calculating the average value from the data as follows:

\[ C_{kj} = \frac{x_{1kj} + x_{2kj} + \cdots + x_{akj}}{a}, \quad j = 1,2,...,p \]

where \( C_{kj} \): centroid of kth-cluster with jth-variable

\( a \): number of members in kth-cluster

3. Step 2 is repeated until there are no more moving objects.

Weatherill & Burton (2009) state that the assessment of the outcome of a partition is an important consideration in cluster analysis. The measurement of cluster quality (for known \( K \)) is the total within-cluster sum of squares (TWCSS). TWCSS formula is as follows:

\[ TWCSS = \sum_{i=1}^{N} \sum_{k=1}^{K} I(x_i \in C_k) \| x_i - m_k \|^2 \]

Where \( m_k \) is the \( C_k \)'s cluster mean and \( I(X) \) is 1 if \( X \) statement is true and 0 if otherwise.

If the data is in the circle unit, the circular distance between two objects is given by \( \cos(\alpha_1 - \alpha_2) \), where \( \alpha_1 \) and \( \alpha_2 \) are the angle [2]. In general, the circular distance to the hypersphere unit, the cosine similarity between two vector units \( y_1 \) and \( y_2 \), is defined as the inner product of \( y_1 \) and \( y_2 \) denoted by \( \langle y_1, y_2 \rangle \). Suppose there are \( n \) spherical data points that will be grouped into \( K \) clusters. The spherical K-means algorithm minimizes:

\[
\sum_{k=1}^{K} \sum_{i=1}^{n} \mu_{ki} \langle y_i, p_k \rangle = \sum_{k=1}^{K} \sum_{i=1}^{n} \mu_{ki} \left( 1 - \frac{\langle y_i, p_k \rangle}{\| y_i \| \| p_k \|} \right) \\
= \sum_{k=1}^{K} \left( \sum_{i=1}^{n} \mu_{ki} \right) - \sum_{i=1}^{n} \left( \mu_{ki} \frac{y_i}{\| y_i \|} \frac{p_k}{\| p_k \|} \right)
\]

For all clusters \( c(i) \in \{1,2,...,k\} \) and for all centroid \( p_k \), where

\[ \mu_{ki} = \begin{cases} 1, & \text{if } c(i) = j \\ 0, & \text{otherwise} \end{cases} \]

The optimization process is an iteration process that consists of the process of calculating the optimal cluster members and cluster centroids, the optimal cluster members for cluster centroids, and the optimal cluster centroids for cluster members. The spherical K-means steps are as follows:

1. Determine the number of clusters, \( K \), and each centroid.
2. Objects are defined as members of the cluster based on their proximity or similarity to the centroid. This step determines the optimal cluster members based on the centroid that was determined in the first step.
3. Recalculate the centroid and update the cluster members based on the resulting centroid. This step determines the optimal centroid for each cluster member.

Perform iteration (repetition) steps 2 and 3 to obtain convergent results [8].
3. Method
The object of this study is the earthquake that occurred in Bengkulu Province from January 1970 to August 2019. The variables are latitude, longitude, magnitude, and depth of the earthquake. The first step in this research is data exploration. Earthquake data exploration aims to describe earthquake events in Bengkulu Province based on the variables used in this research. Thus, we get a summary of the earthquake occurrence in Bengkulu Province. Furthermore, the grouping is carried out using spherical K-means cluster analysis and describing the characteristics of the clusters.

4. Result and discussion

4.1 Descriptive Statistic of Earthquake in Bengkulu Province and Surrounding Areas
The earthquake data used in this study is obtained from the USGS website with earthquake events that occurred from January 1, 1970 to August 17, 2019, in Bengkulu Province. There are 1931 earthquake events consisting of latitude, longitude, depth, and earthquake magnitude. The four variables can be illustrated (using R) in the following figure:

![Earthquake events in Bengkulu Province from 1 January 1970 until 17 August 2019](image)

Figure 1. Earthquake events in Bengkulu Province from 1 January 1970 until 17 August 2019
Based on Figure 1, it can be seen that most of the earthquakes in Bengkulu Province occurred in the sea. An earthquake event with a depth of less than 100 Km (<100) is symbolized by a blue dot, while a red dot is an earthquake event with another depth of less than 300 Km. In other words, shallow earthquakes are mostly centered in the sea. Descriptive statistics of earthquake events in Bengkulu Province can also be seen in the following table:

**Table 1. Descriptive statistic of depth (kilometers)**

| Description    | Value |
|----------------|-------|
| Mean           | 39.84 |
| Standard Deviation | 34.98 |
| Median         | 33    |
| Modus          | 33    |
| Minimum        | 3     |
| Maximum        | 322   |

Based on Table 1, it can be seen that the average depth of the earthquake occurrence in Bengkulu Province is 39.84 Km, with the maximum depth is 322 Km and the minimum depth is 3 Km. Based on this table, it can also be seen that many earthquakes in Bengkulu Province occurred at a depth of 33 km, which is 381 times.

**Table 2. Descriptive statistic of magnitude (Ms)**

| Description    | Value |
|----------------|-------|
| Mean           | 4.94  |
| Standard Deviation | 0.59  |
| Median         | 5     |
| Modus          | 5     |
| Minimum        | 4     |
| Maximum        | 7.9   |

Table 2 describes the descriptive statistic of the earthquake magnitude in Bengkulu Province. Based on the data, the average magnitude of the earthquake occurrence in Bengkulu Province was 4.94 Ms with the smallest value being 4 Ms and the largest value being 7.9 Ms. Bengkulu Province experienced an earthquake with a magnitude of 7.9 Ms in 2007.

4.2 *Application of Spherical K-means Cluster Analysis on earthquake data of Bengkulu province*

Novianti et al. (2017) conducted a K-means cluster analysis on earthquake data in Bengkulu Province using earthquake epicenter and magnitude variables. This study is a development of this research, but the method used is spherical K-means cluster analysis with epicenter, depth, and magnitude variables. The use of spherical K-means is due to the addition of the depth.

The initial stage of the spherical K-means algorithm is to determine the value of K. Rini et al. (2018) determine the optimal number of clusters in earthquake data in Bengkulu Province by using several internal clusters validation indices. The result shows that K = 6 is the optimal number of clusters for earthquake data in Bengkulu Province. Therefore, in this study, the value of K is 6. The results of the spherical K-means cluster analysis with K = 6 can be seen in the following figure:
Figure 2. Results of Spherical K-means Cluster Analysis with K = 6

Based on these 6 clusters, cluster 1 (pink) consists of 55 earthquake points. Cluster 2 (light brown) has 232 earthquake points. Cluster 3 (green) has 172 earthquake points. Cluster 4 (blue-turquoise) consists of 777 earthquake points. Cluster 5 (blue) has 288 earthquake points, while cluster 6 (purple) has 407 earthquake points.

The purpose of using the spherical K-means method in earthquake data in Bengkulu Province is to determine which clusters are formed based on the direction of each earthquake point. Because there is a depth variable that is used as a determinant of the cluster, each earthquake point in a cluster is determined based on similarity in directions or having the same direction (unidirectional). Based on Figure 2, earthquake points that are in a cluster tend to be close together and appear to follow a certain pattern.

4.3 Cluster characteristics with K = 6 for earthquake data in Bengkulu province

Most of the seismic activities in Bengkulu Province are in the sea, which is the Indian Ocean. This is because the position of Bengkulu Province is located in the coastal area, which is traversed by the passage of the Indo-Australian and Eurasian tectonic plates. In addition, tectonic earthquakes also occurred on land that comes from the Semangko fault that stretched across the middle of Sumatra Island. This seismic activity clustering with the spherical K-Means method is grouped into 6 clusters. Based on spherical K-Means cluster analysis, the following table of characteristics is generated:
Table 3 Cluster Characteristic with K = 6

| Cluster | Latitude Average | Longitude Average | Depth Average | Magnitude Average |
|---------|------------------|-------------------|--------------|------------------|
| 1       | -3.38            | 102.87            | 183.98       | 4.89             |
| 2       | -3.89            | 101.38            | 20.73        | 4.90             |
| 3       | -4.07            | 102.56            | 92.00        | 5.09             |
| 4       | -4.34            | 101.73            | 32.84        | 4.94             |
| 5       | -4.16            | 102.11            | 56.99        | 5.09             |
| 6       | -4.12            | 101.42            | 10.47        | 4.80             |

Cluster 1 has an average latitude of -3.38 and an average longitude of 102.87 (mainland Rejang Lebong Regency and its surroundings). Cluster 1 has 55 earthquake points that are the smallest cluster member, an average magnitude of 4.89 Ms, and an average depth of 183.98 km. In this zone, earthquakes rarely occur; this can be seen from the small number of earthquakes. The most earthquake occurrences in this zone were five times in 2011.

Cluster 2 has an average latitude of -3.89 and an average longitude of 101.38 (Ketahun Sea, North Bengkulu, and its surroundings). This zone has an average magnitude of 4.90 Ms, an average depth of 20.73 Km, and 232 earthquake points. Earthquakes that occur in this zone are classified as shallow earthquakes and a zone where earthquakes frequently occur because they have quite a lot of members. The earthquake in this zone is in the Ketahun sea, Ipuh sea, Mukomuko sea, and surrounding seas.

Cluster 3 has an average latitude of -4.07 and an average longitude of 102.56 (mainland Seluma and its surroundings). Cluster 3 has 172 earthquake points, an average magnitude of 5.09 Ms, and an average depth of 92 Km. In this zone, earthquakes occur quite often; this can be seen from a large number of earthquakes. When seen in Figure 2, this zone also includes water areas.

Cluster 4 has an average latitude of -4.34 and an average longitude of 101.73 (Sea of Bengkulu City and its surroundings). This zone has an average magnitude of 4.04 Ms, an average depth of 22.84 Km, and 777 earthquake points. Earthquakes that occur in this zone are classified as shallow earthquakes and a zone where earthquakes are very often happened. It has the most members compared to other zones. Most of the earthquakes in this zone are in the sea of Bengkulu City, South Bengkulu, and its surroundings.

Cluster 5 has an average latitude of -4.16 and an average longitude of 102.11 (South Bengkulu Sea and its surrounding). Cluster 5 has 288 earthquake points, an average magnitude of 5.09 Ms, and an average depth of 56.99 Km. In this zone, earthquakes occur quite often; this can be seen from a large number of earthquakes. See Figure 1; this zone is mostly in the sea area, Enggano Island, and its surroundings.

Cluster 6 has an average latitude of -4.12 and an average longitude of 101.42 (Mentawai and surrounding waters). This zone has an average magnitude of 4.80 Ms, an average depth of 10.47, and 407 earthquake points. Earthquakes that occur in this zone are classified as shallow earthquakes and a zone where earthquakes are very often happened. This zone has the second largest number of members compared to other zones. Most of the earthquakes in this zone are in the sea.
5. Conclusion
During the period 1970-2019, earthquakes in Bengkulu Province and its surroundings mostly occurred in the Indian Ocean, and only a small proportion occurred on land. Bengkulu Province and its surroundings have six earthquake clusters described by the Spherical K-Means method. The result shows that five earthquake zones are in the sea of Bengkulu Province and its surroundings. Another earthquake zone is in the mainland of Bengkulu Province and its surroundings.

6. References
[1] Novianti P, Setyorini D, Rafflesia U 2017 K-Means Cluster Analysis in Earthquake Epicenter Clustering International Journal of Advances in Intelligent Informatics 3 (2) 81-89
[2] Jammalamadaka and SenGupta 2001 Topics in Circular Statistics (Singapore: World Scientific)
[3] Kim S and SenGupta A 2018 Clustering Methods for Spherical Data: an Overview and a New Generalization DOI - 10.1007/978-981-10-8168-2_11
[4] Weatherill G and Burton P W 2009 Delineation of Shallow Seismic Source Zones Using k-means Cluster Analysis, with Application to the Aegean Region Geophysical Journal International 176 565-588.
[5] Hashemi M and Karimi H A Seismic Source Modeling by Clustering Earthquakes and Predicting Earthquake Magnitudes In: Leon-Garcia A. et al (eds) SmartCity 360º, SmartCity 2016, SmartCity 360 2015. Lecture Notes of the Institute for Computer sciences, Social Informatics and Telecommunication Engineering 166 (Springer, Cham)
[6] D S Rini et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 335 012048
[7] Johnson R A and Wichern D W 2002 Applied Multivariate Statistical Analysis Fifth Edition (USA: Pearson Education Inc)
[8] Hornik K, Feinerer I, Kober M, Buchta C 2012 Spherical K-means Clustering Journal of Statistical Software 50 1-22