Desimal: Jurnal Matematika Vol 7 No 3 (2024) 529-540

Contents lists available at DJM

DESIMAL: JURNAL MATEMATIKA

p-ISSN: 2613-9073 (print), e-ISSN: 2613-9081 (online), DOI 10.24042/djm

http://ejournal.radenintan.ac.id/index.php/desimal/index

Cluster analysis of Sumatra Island earthquake distribution

Dian Anggraini[1,2], Sapto Wahyu Indratno[1,3,*], Utriweni Mukhaiyar[1], Antonius Nanang T. Puspito[1,4]

[1] Institut Teknologi Bandung, Indonesia
[2] Institut Teknologi Sumatera, Indonesia
[3] University Center Excellence on Artificial Intelligence for Vision (U-CoE AI-VLB), Indonesia
[4] Center of Earthquake Sience and Technology (CEST), Indonesia

ARTICLE INFO

Article History
Received : 12-10-2024
Revised : 14-11-2024
Accepted : 22-12-2024
Published : 30-12-2024

Keywords:
Cluster Analysis; Earthquake; K-Means Clustering; Risk Clasification.

*Correspondence: E-mail: sapto@itb.ac.id

Doi: 10.24042/djm.v7i3.25108

ABSTRACT

Sumatra Island is highly vulnerable to earthquakes due to multiple seismic sources, including megathrusts, faults, and volcanic activities spanning from Aceh to Lampung. The International Seismological Centre (ISC) has recorded 9,414 earthquakes in Sumatra with magnitudes ranging from 4.0 to 9.1 since 1907. Insufficient preparedness in responding to sudden earthquakes challenges local and central governments in managing impacts. To address this, a risk classification of earthquake-prone areas was conducted using cluster analysis. The "K-means cluster" method identified five earthquake clusters in Sumatra. Cluster 4 has the most events (3,787) but with generally lower magnitudes, resulting in minimal damage. Cluster 2, however, is more concerning due to shallow earthquakes from subduction zones, faults, and volcanoes. This clustering analysis provides critical information for government planning in earthquake risk mitigation and preparedness.

http://ejournal.radenintan.ac.id/index.php/desimal/index

INTRODUCTION

Sumatra Island, the third largest in Indonesia, faces significant natural disaster risks due to its tectonic configuration (Aldi et al., 2023). Located at the intersection of the Eurasian and Indo-Australian plates, the island features an active subduction zone formed by tectonic interactions (Triyoso, 2023). This geological setting gives rise to both tectonic and volcanic earthquake sources, contributing to frequent seismic activity and the potential for major earthquakes.

According to the International Seismological Centre (2024), 9,414 earthquakes with magnitudes ranging from 4.0 to 9.1 have been recorded in Sumatra since 1907. Figure 1 illustrates this distribution.

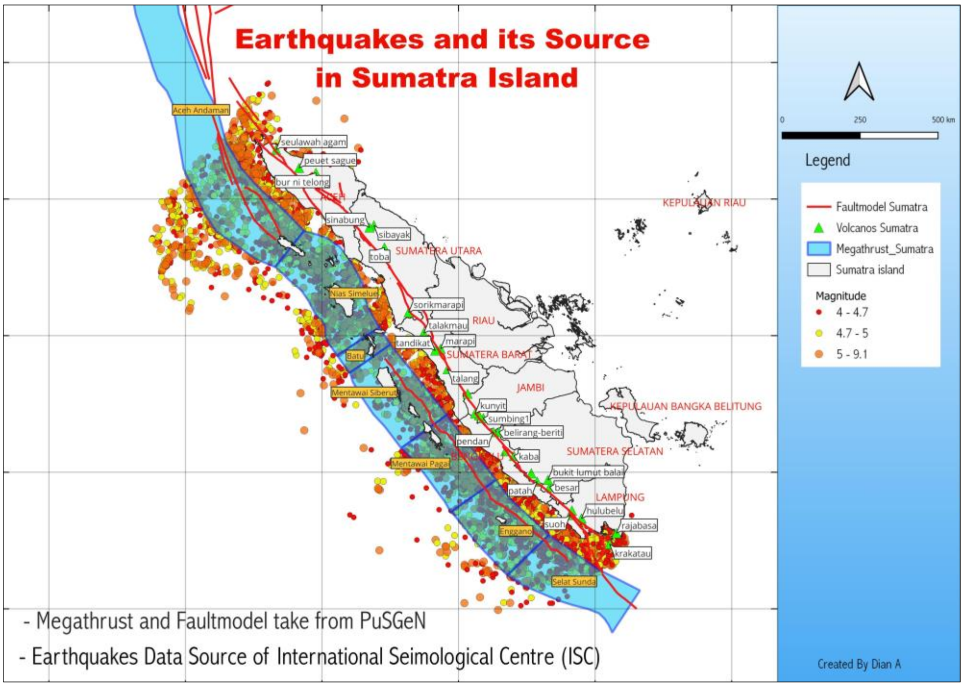

Figure 1. Distribution of Earthquake Occurrences and Sources

Figure 1 illustrates the distribution of earthquake events on Sumatra Island, highlighting megathrust zones, faults, and volcanic sources. Most earthquakes occur on land and offshore in the island's western region.

Earthquakes significantly threaten Sumatra Island due to their unpredictable and unique nature. Even in the same region, each earthquake exhibits distinct characteristics, which complicates identification and prediction efforts (Triyoso, Kongko, Prasetya, & Suwondo, 2024). This variability presents challenges for effective mitigation, as an accurate understanding of the characteristics of past earthquakes is essential for developing successful mitigation strategies (Johnson, Haagenson, Liel, & Rajaram, 2021). Consequently, analyzing the seismicity and the underlying sources of earthquakes, particularly in seismically active regions like Sumatra, is critical to reducing the risk posed by future earthquakes.

The continuous advancement in technology and the growing availability of seismic data, the collection, processing, and analysis of earthquake information have become more efficient and precise. The improvements allow researchers to extract meaningful insights into seismic activity and associated risks. Data mining methods are particularly valuable in this context, providing tools to transform vast and complex datasets into structured, actionable information (Ha, Kambe, & Pe, 2011). These methods encompass a range of tasks, such as estimation, prediction, classification, clustering, and association, each contributing to the understanding of various phenomena (Shah, Shah, Sawant, & Parolia, 2023). Among these, clustering has emerged as a critical approach in the analysis of seismic data due to its ability to group data points with similar characteristics, facilitating the identification of patterns and trends (Annas & Wahab, 2023; Liu, Wang, Zhang, Wei, & Zhou, 2022).

In the context of earthquake studies, clustering provides a robust framework for identifying regions most vulnerable to seismic hazards by categorizing areas based on the frequency and intensity of earthquake events (Yuan, 2021). This method is instrumental in highlighting areas with the greatest potential for damage, offering critical insight for disaster preparedness and mitigation planning (Silitonga et al., 2023). Furthermore, by revealing the spatial distribution of seismic hazards, clustering supports more effective risk management strategies, helping stakeholders prioritize resources and interventions in the most at-risk regions. Through this approach, researchers can better address the challenges posed by seismic activity and enhance societal resilience against earthquakes.

This study employs the k-means clustering algorithm, a commonly used non-hierarchical clustering method, to classify earthquake events on Sumatra Island based on frequency and magnitude. After obtaining the clustering results, a cross-sectional view will be provided to analyze the depth associated with each cluster. The novelty of this study lies in the incorporation of a cross-sectional analysis of depths within the clustering results, offering a unique perspective on the vertical distribution of seismic events in each cluster. K-means clustering is favored for its simplicity and efficiency in processing large datasets, although it requires pre-determination of the number of clusters, which can influence the results (Novianti, Setyorini, & Rafflesia, 2017). Previous studies, such as those by Kertanah et al. (2022), Martadiputra, Rachmatin, & Hidayat (2021), and Novianti et al. (2017), have utilized k-means clustering primarily based on earthquake magnitudes without detailed consideration of source characteristics. In this study, however, the depth of earthquake events is also considered, allowing each cluster to be categorized as deep, medium, or shallow.

This article begins with an introduction to the k-means clustering method and the insight it offers in understanding earthquake patterns. It is followed by a case study on the application of k-means clustering to analyze the distribution of earthquake events on Sumatra Island. The final section provides a conclusion to understanding seismic risks and supporting risk mitigation efforts. The results offer valuable data for government planning, spanning policy development, technical measures, and financing in earthquake mitigation and prevention.

K-means cluster

K-means is a commonly applied method for grouping observed objects into clusters (Annas & Wahab, 2023). This approach partitions objects into one or more clusters based on their proximity to each other, such that objects with the shortest distances between them are grouped and share similar characteristics (Sartika, Murniati, Binarto, & Habinuddin, 2022).

Steps for Using K-means Clustering:

1. Determine the Number of Clusters: In K-means clustering, selecting the initial number of clusters is crucial. Common methods for determining this include:
   - Elbow Method: This approach identifies the optimal number of clusters by observing the point on a graph where the rate of decrease in error slows, creating an “elbow” shape (Shi et al., 2021).
   - Silhouette Method: The silhouette score assesses cluster quality, indicating whether the objects are appropriately grouped. The steps for calculating the silhouette score are as follows (Kariyavula & Anbarasan, 2023):
     a. Calculate the average distance of an object i to all other objects in its cluster, denoted as $a(i)$.
     b. Calculate the average distance of object i to all objects in other clusters, denoted as $b(i)$, and take the minimum value.
     c. The silhouette coefficient $S(i)$ is calculated as:

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

where:

$$S(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & ; if\ a(i) < b(i) \\ 0 & ; if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & ; if\ a(i) > b(i) \end{cases}$$

Silhouette values range from -1 to 1. Values close to 1 indicate well-clustered objects, while values near 0 suggest boundary cases and values around -1 indicate incorrect clustering. The optimal cluster count corresponds to the highest average silhouette score among clusters.

2. Select Initial Cluster Centers (Centroids) Randomly: Define K initial centroids, selecting them randomly.
3. Calculate the Distance of Each Object to the Cluster Center: Assign each object to the nearest centroid, often using the Euclidean distance, calculated as:

$$d(x_i, x_j) = \sqrt{\left(x_i - x_j\right)^2} = \sqrt{\textstyle\sum_{m=1}^{n}\left(x_{im} - x_{jm}\right)^2}$$

where:
$d(x_i, x_j)$: Euclidean distance
x_i : data- i
x_j : data- j
x_{im} : data- i attribute – m
x_{jm} : data- j attribute – m

4. Update the Cluster Centers: Calculate the new center of each cluster by averaging the positions of all objects within the cluster.
5. Recalculate Distances and Reassign Objects: Recompute the distance between each object and the updated cluster centers. If no changes occur in cluster assignments, the clustering process is complete. If changes occur, return to step 4 and repeat until all cluster assignments stabilize.

Earthquake

The concept of earthquakes was first introduced in 1910 by Rheid, following observations of seismic activity along the San Andreas Fault in California (Mohadjer et al., 2021). This concept, known as the Elastic Rebound Theory, suggests that the earth’s crust behaves as an elastic material that gradually deforms under accumulating stress. When this stress surpasses the rock's elastic limit, the crust experiences a sudden, irreversible shift, releasing accumulated energy as elastic waves.

In the context of earthquakes, these elastic waves are referred to as seismic waves, which can be classified into body waves and surface waves (Ünal, Askan, & Selcuk-Kestel, 2017). Body waves, which travel through the interior of the earth, originate directly from the earthquake's focal point. They include two types: P-waves (primary or compression waves) and S-waves (secondary or shear waves). Surface waves, on the other hand, propagate along the earth's surface, resulting from the interaction of body waves with free surfaces and shallow structures. The main types of surface waves are Rayleigh waves and Love waves (Rizal, Yodi Gunawan, W. Indratno, & Meilano, 2023). Table 1 presents a classification of earthquakes based on their depth and associated effects.

Table 1. Classification of Earthquakes by Depth Level (source: Badan Penanggulangan Bencana Daerah (2018))

Type of Earthquake	Depth Level (km)	Effect
Shallow earthquakes	≤ 60	Cause significant damage and can cause tsunamis
Moderate earthquake	$60 - 300$	Inflicts light damage and vibrations more pronounced
Earthquake deep	≥ 300	Not dangerous

METHOD

This study utilizes earthquake data recorded on Sumatra Island, with magnitudes ranging from 4.0 to 9.1, spanning 115 years from 1907 to 2022. The data for this case study was obtained from the International Seismological Center (ISC) global seismic catalogs, which provide real-time earthquake data. Four variables are analyzed: longitude, latitude, magnitude, and depth.

Data analysis and visualization were conducted using R software and QGIS for mapping. In R, data processing involved the use of the 'factoextra,' 'cluster,' 'tidyverse,' and 'dplyr' libraries. The following flowchart illustrates the k-means clustering process.

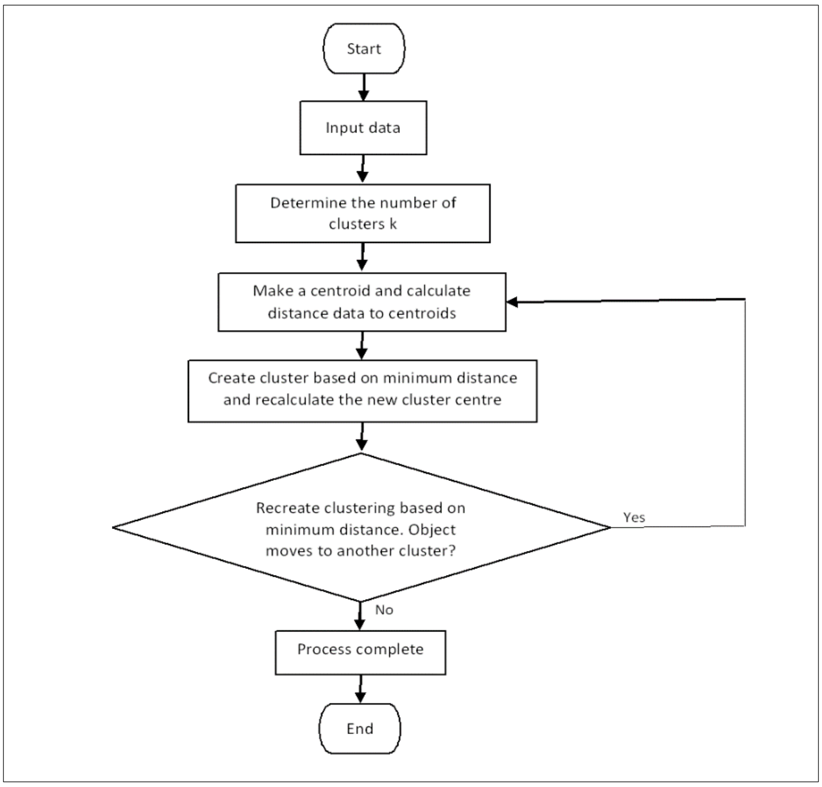

Figure 2. K-means Clustering Process

RESULTS AND DISCUSSION

Earthquakes can be categorized into three main types based on their causes: tectonic, volcanic, and collapse earthquakes. Tectonic earthquakes result from shifts in the earth's layers due to energy release in subduction zones. These shifts produce subduction pathways and fault lines, where shallow earthquakes (less than 60 km deep) often occur along fault lines, while subduction zones host earthquakes that can range from shallow to deep. Volcanic earthquakes are associated with volcanic eruptions. Figure 3 illustrates the distribution of these earthquake sources on Sumatra Island.

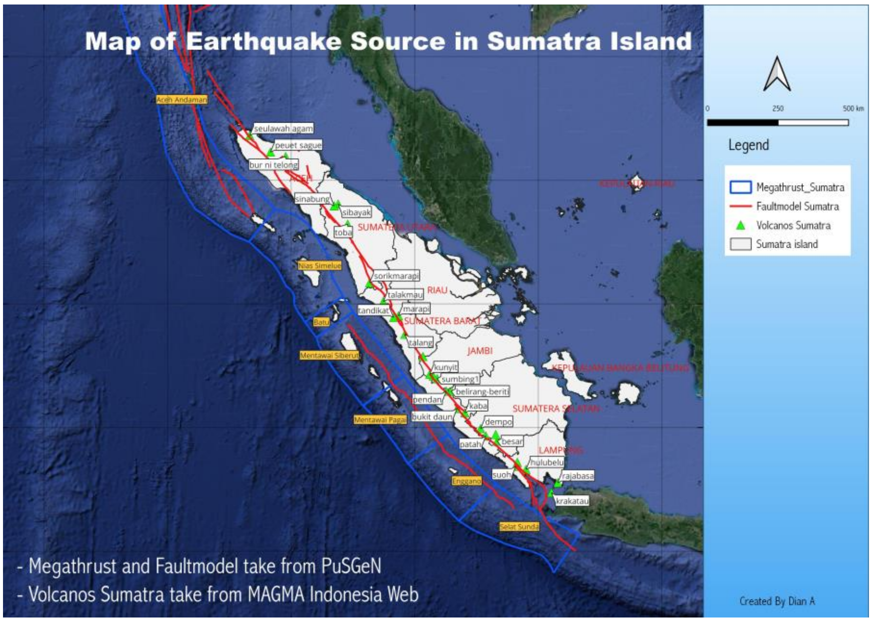

Figure 3. Distribution of Earthquake Sources in Sumatra

Figure 3 shows three main sources of earthquakes in Sumatra. Along the western coast, seven subduction zones run from north to south, including Aceh-Andaman, Nias-Simelue, Batu, Mentawai-Siberut, Mentawai-Pagai, Enggano, and the Sunda Straits. These zones are formed by the convergence of the Indian Ocean and Eurasian plates beneath Sumatra. Additionally, fault lines serve as critical earthquake sources; faults are fractures in the rock where significant displacement has occurred. The red lines in Figure 3 indicate fault lines extending longitudinally across Sumatra. Active volcanoes, marked by green triangles, are also significant earthquake sources, following the "Ring of Fire." Both fault lines and volcanoes are found along the same general path.

Figure 1 and Table 2 provide an overview of earthquake data on Sumatra Island from 1907 to 2022, sourced from the ISC, with magnitudes ranging from 4.0 to 9.1. This data reveals Sumatra's high vulnerability to earthquakes, with events occurring from the island's northern to southern regions. Most earthquakes are shallow, which is significant because shallow earthquakes typically cause substantial damage. Based on Figure 4, 58% of recorded earthquakes occurred in subduction areas, 14% on land, and the remaining 28% offshore.

Table 2. Summary of Earthquake Magnitude and Depth in Sumatra

Variables	Min	Average	Max
Magnitude	4.0	4.886	9.1
Depth (km)	0	43.08	582

To assess earthquake risk in Sumatra, a clustering analysis was conducted based on magnitude and depth. The goal of clustering is to identify areas with higher potential for damage according to earthquake magnitude and focal depth. Using the k-means algorithm, data were grouped, with the Elbow and Silhouette methods applied to determine the optimal number of clusters for earthquake magnitude data.

Figures 4 and 5 show the Elbow and Silhouette methods results, suggesting different optimal cluster counts. The

Elbow method indicates an optimal number of clusters at $K = 5$, while the Silhouette method suggests $K = 4$. To resolve this, the GAP statistics were used as an additional method, as shown in Figure 6.

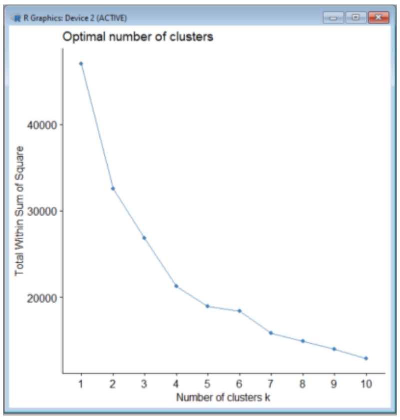

Figure 4. Optimal Number of Clusters Determined by the Elbow Method

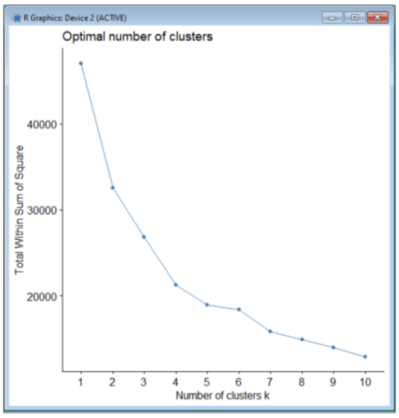

Figure 5. Optimal Number of Clusters Determined by the Silhouette method

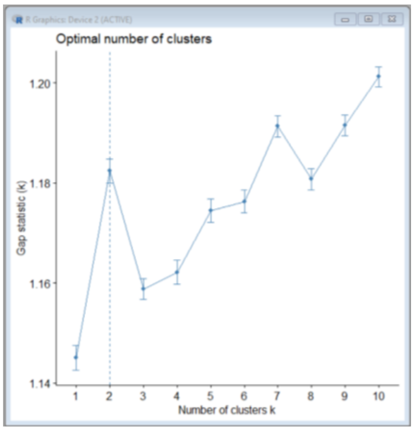

Figure 6. Optimal Number of Clusters Determined by the GAP Statistic

Figure 6 confirms that the optimal number of clusters is $K = 5$. Therefore, $K = 5$ was chosen as the initial setting for k-means clustering. The clustering process was then applied, grouping the data into five distinct clusters. The results are summarized in Table 3.

The distribution of earthquakes and their depths across the five clustered regions is shown in Figures 7 until 11.

Table 3. Cluster Summary

Category		Group				
		1	2	3	4	5
Magnitude	min	4.6	4	4	4	4.6
	means	5.232	4.795	4.636	4.751	5.6
	max	7.6	6.3	7.3	5.6	9.1
Depth (km)	min	6	0	92	0	3
	means	48.23	35.12	159.5	32.06	34.26
	max	179	117	582	115	132
N		1027	3189	593	3787	818

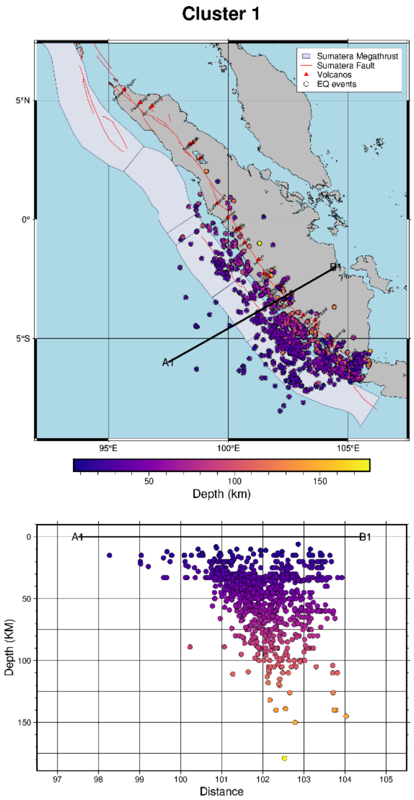

Figure 7. Cluster 1

Cluster 1 (Figure 7) recorded 1,027 earthquake events, with magnitudes ranging from 4.6 to 7.6 and an average magnitude of 5.23, categorizing them as significant and potentially damaging earthquakes. The map shows that these events are primarily distributed in the western region, from North Sumatra to the Sunda Strait, with higher concentrations in South Sumatra to Lampung, particularly around the Enggano subduction zone.

The earthquakes in Cluster 1 are largely shallow, with depths less than 100 km below sea level, indicating their source in the subduction area.

With a substantial number of occurrences and an average magnitude of 5.23, Cluster 1 is considered vulnerable to earthquake damage due to the predominance of shallow events.

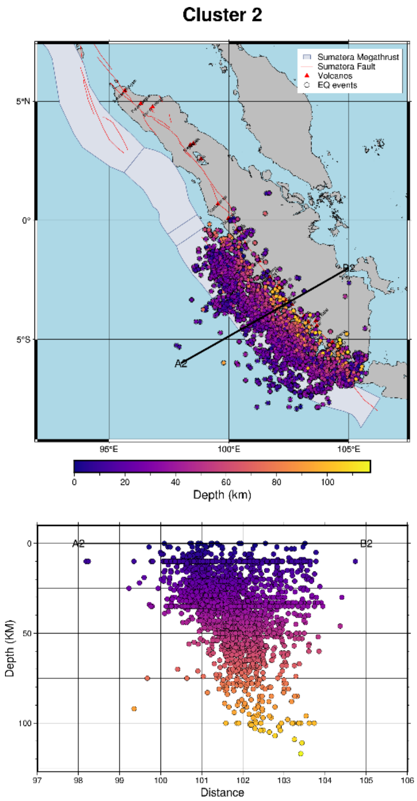

Figure 8. Cluster 2

Figure 8 shows that earthquake events in Cluster 2 are distributed from West Sumatra to the Sunda Strait, in an area similar to Cluster 1 but with a higher concentration around the Mentawai Pagai and Enggano subduction zones. Cluster 2 recorded 3,189 events, with magnitudes ranging from 4.0 to 6.3 and an average magnitude of 4.8. The majority of these earthquakes are classified as shallow or moderate, with depths of less than 100 km.

Additionally, the presence of Mount Anak Krakatau in this area adds potential for seismic activity. Compared to Cluster 1, Cluster 2 has a higher potential for damage due to the combination of shallow earthquake events and the additional seismic influence of volcanic activity.

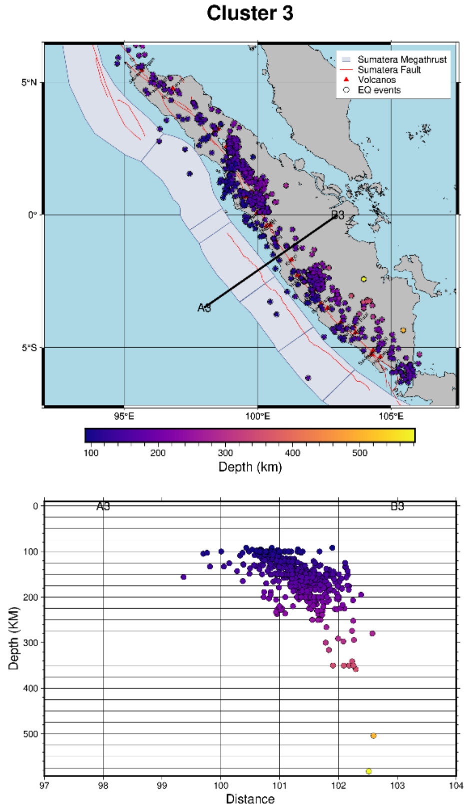

Figure 9. Cluster 3

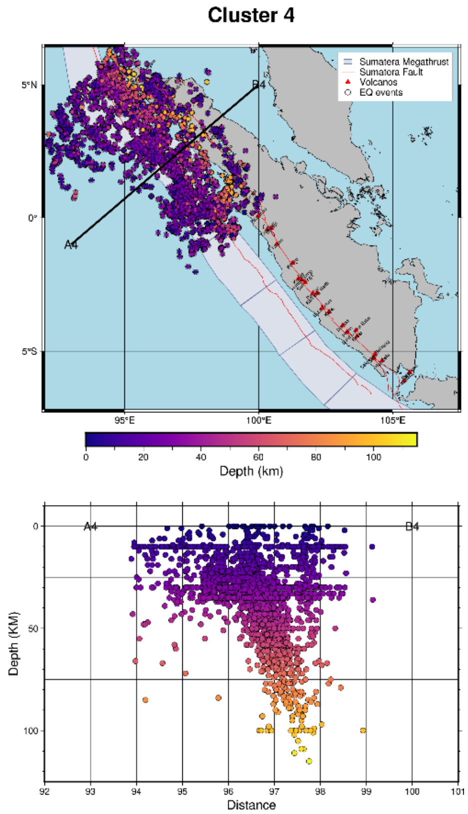

Figure 101. Cluster 4

The distribution of earthquakes in Cluster 3 (Figure 9) is concentrated around fault zones and volcanic chains, stretching from northern Aceh to the southern Sunda Strait. Notably, the areas around the Toba and Sorik Marapi volcanoes serve as focal points for seismic activity in this cluster. Cluster 3 contains fewer events than other clusters, with 593 recorded occurrences and magnitudes ranging from 4.0 to 7.3. Although earthquakes in this region are less frequent, significant events have occurred. Most earthquakes in Cluster 3 are classified as moderate to deep, with depths exceeding 100 km.

Cluster 4 (Figure 10) exhibits a high concentration of earthquakes in the northern region of Sumatra Island, with the majority occurring offshore. This cluster records the highest number of events compared to others, totaling 3,787 with magnitudes ranging from 4.0 to 5.6. Although the magnitude range is relatively narrow, the frequency of occurrences is the highest among all clusters. Most earthquakes in Cluster 4 are shallow, with a few classified as moderate in depth.

The vulnerability level for Cluster 4 remains manageable, as the area is largely distant from human settlements. However, the frequent seismic activity in this region may generate sea waves due to offshore vibrations.

Earthquake occurrences in Cluster 5 (Figure 11) are similar to those in Cluster 4, though with a distribution that extends further south. This cluster includes seismic activity from three primary sources: subduction zones, fault lines, and volcanic areas. A total of 818 events were recorded, with magnitudes ranging from 4.6 to 9.1 and an average magnitude of 5.6. In terms of depth, most earthquakes in Cluster 5 are shallow, with some classified as moderate.

The prevalence of shallow earthquakes in this region raises

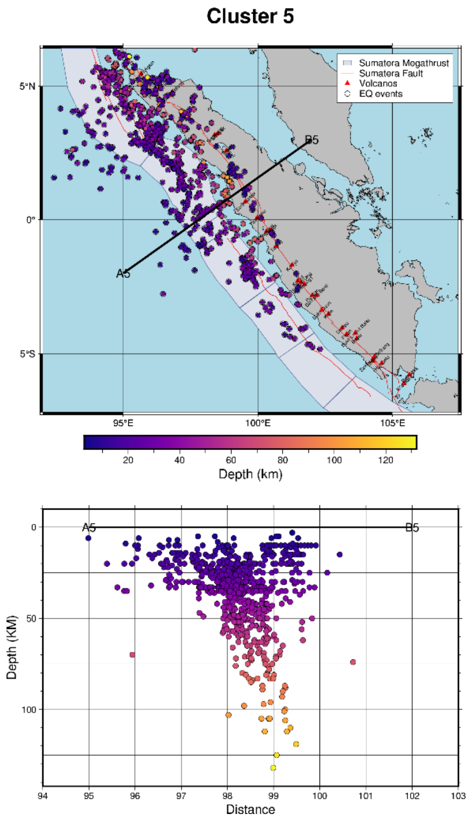

Figure 11. Cluster 5

concern, as such events can be highly destructive. This was evident in the 2004 Aceh earthquake and tsunami, which struck this area with a magnitude of 9.1, resulting in devastating effects.

CONCLUSIONS AND SUGGESTIONS

The clustering analysis using the k-means algorithm on 9,414 earthquake events recorded on Sumatra Island identified five distinct clusters. The findings highlight that 58% of earthquakes occur in subduction zones, 14% on land, and 28% in offshore areas beyond the subduction and mainland zones. Among these, Cluster 4 contains the highest number of events (3,787), but the limited range of magnitudes in this cluster indicates minimal potential for damage. In contrast, Cluster 2 stands out as the most concerning due to its predominance of shallow earthquakes originating from subduction zones, faults, and active volcanoes. This combination of seismic sources in Cluster 2 suggests a significantly higher potential for damage compared to other clusters.

Although earthquake prediction remains beyond current scientific capability, this study provides valuable insights to support disaster mitigation strategies in earthquake-prone regions. It is recommended that future research further refine the analysis by differentiating areas based on specific seismic sources, such as subduction zones, faults, and volcanic activities. Such efforts would enhance the understanding of spatial earthquake distributions and contribute to more targeted mitigation and preparedness measures for Sumatra Island.

REFERENCES

Aldi, M., Silaen, N., Siregar, I. R., Dhaifullah, D., Suhardja, S. K., & Madrinovella, I. (2023). Processing receiver function data in mentawai region, sumatra. *IOP Conference Series: Earth and Environmental Science*, *1227*(1). https://doi.org/10.1088/1755-1315/1227/1/012047

Annas, M., & Wahab, S. N. (2023). Data mining methods: K-means clustering algorithms. *International Journal of*

Cyber and IT Service Management, *3*(1).

Badan Penanggulangan Bencana Daerah. (2018). Pengertian gempa bumi, jenis-jenis, penyebab, akibat, dan cara menghadapi gempa bumi. *Https://Bpbd.Bandaacehkota.Go.Id*. Retrieved from https://bpbd.bandaacehkota.go.id/2018/08/05/pengertian-gempa-bumi-jenis-jenis-penyebab-akibat-dan-cara-menghadapi-gempa-bumi/

Ha, J., Kambe, M., & Pe, J. (2011). Data mining: Concepts and techniques. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

International Seismological Centre. (2024). Earthquake data for sumatra island. Retrieved December 14, 2024, from https://www.isc.ac.uk/

Johnson, E. G., Haagenson, R., Liel, A. B., & Rajaram, H. (2021). Mitigating injection-induced seismicity to reduce seismic risk. *Earthquake Spectra*, *37*(4). https://doi.org/10.1177/87552930211008479

Kariyavula, S., & Anbarasan, K. (2023). *Improved accuracy of cluster formation for customer segmentation by identification of novel optimal number of clusters using silhouette method over k-means clustering*. 020100. https://doi.org/10.1063/5.0119477

Kertanah, K., Rahadi, I., Aryani Novianti, B., Syahidi, K., Sapiruddin, S., Mandala Putra, H., … Sabar, S. (2022). Applying k-means algorithm for clustering analysis earthquakes data in west nusa tenggara province. *Indonesian Physical Review*, *5*(3), 197–207. https://doi.org/10.29303/ipr.v5i3.148

Liu, J., Wang, Y., Zhang, Q., Wei, J., & Zhou, H. (2022). Scientometric analysis of public health emergencies: 1994–2020. *International Journal of Environmental Research and Public Health*, *19*(2), 640. https://doi.org/10.3390/ijerph19020640

Martadiputra, B. A. P., Rachmatin, D., & Hidayat, A. S. (2021). Analysis of characteristics of earthquake area in indonesia in 2020 with cluster analysis as natural disaster. *International Journal of Science and Research (IJSR)*, *9*(11).

Mohadjer, S., Mutz, S. G., Kemp, M., Gill, S. J., Ischuk, A., & Ehlers, T. A. (2021). Using paired teaching for earthquake education in schools. *Geoscience Communication*, *4*(2), 281–295. https://doi.org/10.5194/gc-4-281-2021

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. https://doi.org/10.26555/ijain.v3i2.100

Rizal, J., Yodi Gunawan, A., W. Indratno, S., & Meilano, I. (2023). Seismic activity analysis of five major earthquake source segments in the sumatra megathrust zone. *Bulletin of the New Zealand Society for Earthquake Engineering*, *56*(2), 55–70. https://doi.org/10.5459/bnzsee.1555

Sartika, E., Murniati, S., Binarto, A., & Habinuddin, E. (2022). Penerapan k-means cluster dan evaluasi clustering pada pesebaran kasus covid-19. *STATISTIKA Journal of Theoretical Statistics and Its Applications*, *22*(2), 147–156. https://doi.org/10.29313/statistika.v22i2.1229

Shah, K., Shah, N., Sawant, V., & Parolia, N. (2023). *Practical data mining techniques and applications*. Boca Raton: Auerbach Publications.

https://doi.org/10.1201/9781003390220

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2021*(1), 31. https://doi.org/10.1186/s13638-021-01910-w

Silitonga, B. E., Suardi, I., Firmansyah, A., Hanif, M., Ramdhan, M., & Sembiring, A. S. (2023). Tectonic structure of northern sumatra region based on seismic tomography of p and s wave velocity. *EKSPLORIUM*, *44*(1), 1. https://doi.org/10.55981/eksplorium.2023.6784

Triyoso, W. (2023). Probabilistic seismic hazard function based on spatiotemporal earthquake likelihood simulation and akaike information criterion: The pshf study around off the west coast of Sumatra Island before large earthquake events. *Frontiers in Earth Science*, *11*. https://doi.org/10.3389/feart.2023.1104717

Triyoso, W., Kongko, W., Prasetya, G. S., & Suwondo, A. (2024). Study on earthquake and tsunami hazard: Evaluating probabilistic seismic hazard function (pshf) and potential tsunami height simulation in the coastal cities of sumatra island. *Frontiers in Built Environment*, *10*. https://doi.org/10.3389/fbuil.2024.1310251

Ünal, B., Askan, A., & Selcuk-Kestel, A. S. (2017). Simulation of large earthquakes and its implications on earthquake insurance rates: A case study in bursa region (turkey). *Natural Hazards*, *85*(1), 215–236. https://doi.org/10.1007/s11069-016-2578-4

Yuan, R. (2021). An improved k-means clustering algorithm for global earthquake catalogs and earthquake magnitude prediction. *Journal of Seismology*, *25*(3), 1005–1020. https://doi.org/10.1007/s10950-021-09999-8