Earthquake Clustering in Sumatra Using K-Medoids Based on Magnitude and Depth

Demil Saputra*, Dwi Agustin Nuriani Sirodj

Study Program of Statistics, Faculty of Mathematics and Natural Science, Bandung Islamic University, Indonesia.

10060122003@unisba.ac.id, dwi.agustinnuriani@unisba.ac.id

Abstract. Earthquake activity in Sumatra exhibits high variability in terms of magnitude and focal depth due to its complex tectonic setting, making it difficult to identify general patterns using conventional descriptive analysis. This study aims to explore mainshock earthquake grouping patterns in Sumatra based on magnitude and depth variables as an initial effort to understand the characteristics of seismic events in the region. Earthquake data from January 2000 to October 2025 were obtained from the International Seismological Centre (ISC) and the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG), consisting of 1037 mainshock earthquake events in Sumatra with a minimum magnitude of 5 moment magnitude. Data preprocessing included outlier detection using boxplots, multicollinearity testing, and standardization using Z-score. The clustering process was conducted using the K-Medoids method with Euclidean distance, and the optimal number of clusters was determined based on the Silhouette Coefficient. The results show that the optimal clustering configuration consists of three clusters with a Silhouette Coefficient value of 0.405, indicating a moderate level of cluster separation. Each cluster represents distinct mainshock earthquake characteristics based on magnitude and depth, and their spatial distributions exhibit different patterns across the Sumatra region. These findings indicate that the K-Medoids method is suitable for exploratory clustering of mainshock earthquake data with large variability and the presence of outliers. The results provide an initial overview of mainshock earthquake characteristics in Sumatra and serve as a basis for further research incorporating additional variables or spatio-temporal approaches to improve the understanding of regional seismicity.

Keywords: *Earthquake, K-Medoids, Magnitude.*

*10060122003@unisba.ac.id

A. Introduction

Indonesia is one of the countries with the highest levels of seismic activity in the world. This condition is primarily attributed to its geographical position at the convergence of several major tectonic plates, namely the Indo-Australian Plate, the Eurasian Plate, and the Pacific Plate. As a result, Indonesia, particularly the island of Sumatra, frequently experiences earthquakes with diverse characteristics in terms of both magnitude and focal depth [1].

Sumatra Island is recognized as one of the regions with intense seismic activity due to the presence of an active subduction zone along its western offshore area and the Sumatra Fault System extending from the northern to the southern part of the island [2]. Geographically, Sumatra is located between 95°–105° east longitude and 6° north–6° south latitude [3]. Earthquake occurrences in this region are not only frequent but also exhibit considerable variability in magnitude and depth. Such variability indicates that each earthquake event possesses distinct characteristics and may generate different impacts on the environment and society [4].

In seismic studies, earthquake magnitude and focal depth are two key variables commonly used to describe the characteristics of seismic events. Magnitude represents the amount of energy released during an earthquake, whereas depth indicates how deep the earthquake source is located beneath the Earth's surface [5]. Although both variables are essential, the relationship between magnitude and depth is not always linear or uniform across events [6]. Therefore, an analytical approach capable of capturing underlying data patterns and structures in a comprehensive manner is required.

One approach that can be employed to identify patterns in earthquake occurrences is clustering analysis. Through clustering, earthquake events can be grouped into several clusters based on similarities in their characteristics, providing an overall description of the types of earthquakes occurring in Sumatra [7]. This approach facilitates the simplification of complex earthquake datasets into more interpretable groups.

However, earthquake data, particularly the depth variable, often exhibit uneven distributions and contain extreme values or outliers. Earthquake depths can range from shallow to very deep events, resulting in a wide value range [8]. This condition presents challenges in clustering analysis, especially when using methods that are sensitive to outliers.

Therefore, the K-Medoids method is employed in this study due to its robustness in handling data containing outliers [9]. Unlike the K-Means method, which uses mean values as cluster centers, K-Medoids selects actual data points (medoids) as cluster representatives, making it less sensitive to extreme values [10]. These characteristics make K-Medoids more suitable for clustering earthquake events based on magnitude and depth, which exhibit substantial variability and potential outliers.

Based on this background, this study focuses on clustering mainshock earthquake events in Sumatra Island using magnitude and depth variables through the K-Medoids method. Mainshock events are used to ensure that the analyzed data represent independent seismic occurrences, thereby reducing the influence of dependent events such as aftershocks that may bias the clustering results. The findings are expected to provide clearer insights into the spatial distribution and characteristic patterns of mainshock earthquake occurrences in Sumatra, as well as to contribute to a better understanding of mainshock earthquake characteristics based on the resulting clusters.

B. Method

Data Sources

The data used in this study were obtained from two sources, namely the International Seismological Centre (ISC) and the Meteorology, Climatology, and Geophysics Agency of Indonesia (BMKG). The dataset consists of 1037 mainshock earthquake events in Sumatra recorded during the period from January 2000 to October 2025, with a minimum magnitude threshold of 5.0 M_w (moment magnitude).The dataset includes four variables: longitude (°), latitude (°), magnitude (M_w), and depth (km). Among these variables, depth (X_1) and magnitude (X_2) were selected as the primary variables for the clustering process. The summary of the research data is presented in Table 1.

Table 1. Research Dataset

No.	Longitude (°)	Latitude (°)	Depth (km)	Magnitude (M_w)
1	98,0467	2,0181	43,1	5,1

No.	Longitude (°)	Latitude (°)	Depth (km)	Magnitude (M_w)
2	98,7549	-1,385	25	5,1
3	95,9112	4,7683	31	5,1
⋮	⋮	⋮	⋮	⋮
1035	98,7541	1,2435	92	5,5
1036	96,4281	1,9632	22	5,6
1037	96,0333	4,8728	5	5,3

The research stages are illustrated in the flowchart presented in Figure 1.

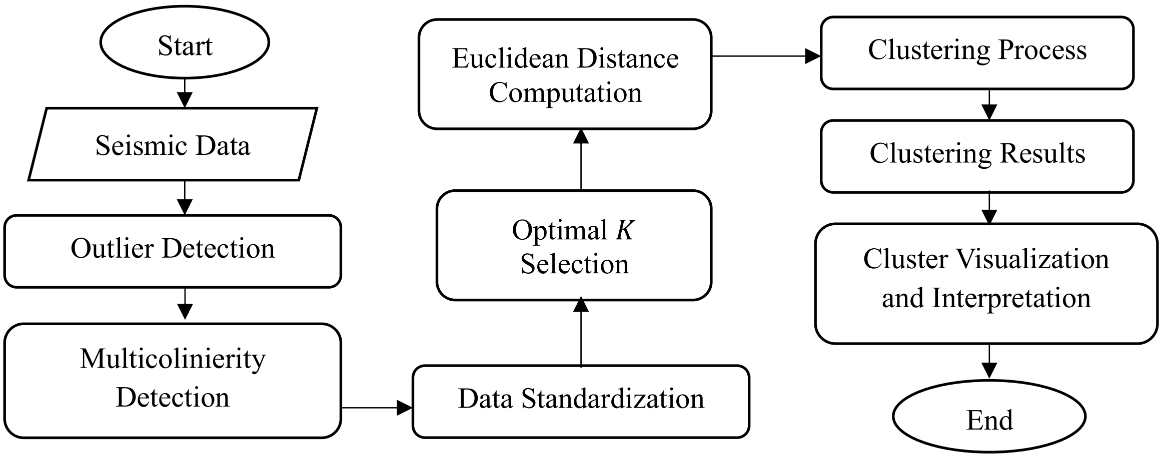


Figure 1. Research Flow Chart

Outlier Detection

Outlier detection in this study was performed using boxplots to identify anomalous observations in each variable [11]. Through boxplots, data distributions are represented by the first quartile ($Q1_m$), the third quartile ($Q3_m$), and the interquartile range (IQR_m) for the m-th variable. Observations located outside the boxplot whiskers are identified as outliers.

Mathematically, the lower and upper bounds used for outlier detection in the boxplot for the m-th variable are defined as follows:

$$\text{Lower bound}_m = Q1_m - 1{,}5 \times IQR_m \quad ...(1)$$

$$\text{Upper bound}_m = Q3_m + 1{,}5 \times IQR_m \quad ...(2)$$

where $m = 1{,}2$ represent the earthquake magnitude and depth variables, respectively.

Multicolinierity Detection

Multicollinearity detection was conducted to ensure that the variables did not exhibit strong linear relationships with one another. This step is essential to ensure that each variable contributes unique information, thereby improving the accuracy and reliability of the clustering results [12]. Multicollinearity was evaluated using the Variance Inflation Factor (VIF), which is calculated as follows:

$$VIF_m = \frac{1}{1-R_m^2} \quad ...(3)$$

where R_m^2represents the coefficient of determination obtained from the regression of the m-th independent variable on the remaining independent variables. A variable is considered to exhibit multicollinearity if the VIF value exceeds 10.

Data Standardization

In this study, the original variables $X_i$were standardized to have a mean of 0 and a standard deviation of 1. This standardization process was applied to eliminate differences in scale among variables, ensuring that each variable contributes equally to the distance calculations [13]. Standardization was performed using the following formula:

$$Z_{im} = \frac{X_{im} - \bar{X}_m}{s_m} \quad ...(4)$$

where Z_{im} denotes the standardized value of the i-th observation for the m-th variable.

Euclidean Distance

In this study, the similarity between earthquake events was measured using the Euclidean distance. This distance metric was employed to calculate the distance between data points based on earthquake magnitude and depth variables in a two-dimensional feature space. The use of Euclidean distance provides a simple and interpretable quantitative measure of proximity between earthquake events [14]. The Euclidean distance is defined as:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2} \quad ...(5)$$

Using Euclidean distance, earthquake events with similar magnitude and depth values have smaller distances and tend to be grouped within the same cluster. This distance metric was selected because it is appropriate for continuous numerical data and is commonly used in distance-based clustering methods, including K-Medoids.

K-Medoids

K-Medoids is a partition-based clustering method that uses representative objects, referred to as medoids, as cluster centers instead of the mean values of objects within a cluster [15]. The medoid of a cluster can be defined as follows:

$$x_{m_l} = \arg\min_{x_i \epsilon C_l} \sum_{x_i \epsilon C_l} d(x_i, x_j) \quad ...(6)$$

x_{m_l} denotes the medoid of the l-th cluster. The medoid is obtained by selecting an observation x_i within cluster C_l that minimizes the total distance to all other observations x_j in the same cluster C_l. Consequently, the medoid of the l-th cluster represents the most central observation, as it has the smallest total distance compared to other observations within the cluster.

The steps of the K-Medoids algorithm are described as follows:

1. Determining the number of clusters $k$to be formed.
2. Selecting $k$objects from the dataset as the initial medoids.
3. Calculating the distance between each object and each medoid, and assigning objects to the nearest cluster.
4. Computing the total distance of all objects to the medoid within each cluster.
5. Selecting another object within the cluster as a candidate new medoid.
6. Recalculating the distances and total distances using the candidate medoid.
7. Comparing the previous total distance with the new total distance; if the new total distance is smaller, the candidate medoid replaces the previous medoid.
8. Repeating the process until no further changes in medoids occur and the final clusters are obtained.

Silhouette Coefficient

The Silhouette Coefficient is used to evaluate the quality of clustering results by measuring how well an object fits within its assigned cluster compared to other clusters [16]. For each object i, the silhouette value is computed based on two measures: $a(i)$, which represents the average distance between object $i$and all other objects within the same cluster, and $b(i)$, which denotes the minimum average distance between object $i$and objects in the nearest neighboring cluster. The Silhouette Coefficient for the i-th object is defined as:

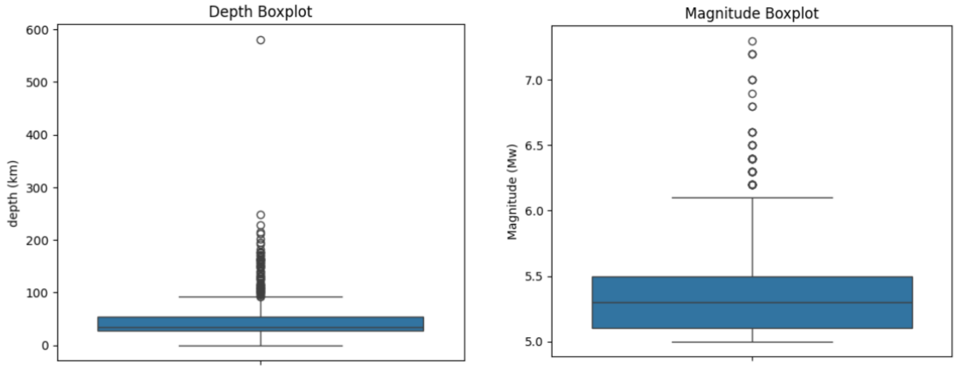
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad ...(7)$$

The silhouette coefficient ranges from -1 to 1, where values close to 1 indicate that the object is well clustered, values close to 0 indicate that the object lies near the boundary between clusters, and negative values suggest potential misclassification. The average silhouette value across all objects is used to assess the overall quality of the clustering results.

C. Result and Discussion

Pre-processing Data

At the initial stage of the analysis, the mainshock earthquake event data were first subjected to outlier detection process. Outlier detection was applied to the magnitude and depth variables using boxplots, as illustrated in Figure 2.



(a) (b)

Figure 2. Depth and Magnitude Boxplot

As shown in Figure 2, both variables exhibit outlier values, indicating the presence of earthquake events with greater depths and larger magnitudes compared to the majority of events. This condition supports the use of the K-Medoids clustering method, which is more suitable for datasets containing outliers in their variables.

The next step involved examining the relationship between the depth and magnitude variables. This assessment was carried out by evaluating the Variance Inflation Factor (VIF) values for both variables, as presented in Table 2.

Table 2. VIF Value of Each Variable

Variable	VIF
Depth (X_1)	1.000683
Magnitude (X_2)	1.000683

Based on Table 2, none of the variables exhibit VIF values greater than 10, indicating that multicollinearity is not present between the variables.

Subsequently, prior to the clustering process, the scales of both variables were standardized. This step was necessary because the depth and magnitude variables are measured in different units and have substantially different value ranges. Therefore, both variables were standardized using the Z-score method. An example of the Z-score calculation for the depth variable of the first earthquake event is provided below.

$$Z_{1,1} = \frac{X_{1,1} - \bar{X}_1}{s_1} = \frac{43.1 - 46.22}{35.1} = -0.089$$

The standardization procedure was applied to all observations, The standardized values of depth (Z_1) and magnitude (Z_2) are presented in Table 3.

Table 3. Dataset After Standardization

No.	$\boldsymbol{Z_1}$	$\boldsymbol{Z_2}$
1	-0.089	-0.690
2	-0.605	-0.690
3	-0.434	-0.690
4	0.441	-0.429
5	0.353	-0.951
⋮	⋮	⋮
1716	1.305	0.353

No.	Z_1	Z_2
1717	-0.690	0.613
1718	-1.175	-0.169

Through this standardization process, both variables generally fall within comparable value ranges, ensuring that the clustering results are not dominated by a particular variable.

Clustering Process

In the K-Medoids procedure, the initial step is determining the number of clusters k. The choice of $k$was based on the Silhouette Coefficient, where the selected $k$corresponds to the highest silhouette value. Table 4 presents the Silhouette Coefficient values for $k = 2, 3, 4, 5, 6$, and 7.

Table 4. Silhouette Coefficient of Each Clusters

k	Silhouette Coefficient
2	0.288
3	0.405
4	0.286
5	0.217
6	0.325
7	0.369

From Table 4, the largest Silhouette value occurs at $k = 3$, namely 0.405. This result indicates that three clusters yield a more optimal grouping compared to the other tested values of k. Consequently, the number of clusters used in this study is three.

The subsequent step is to determine the initial medoids randomly for each cluster. In this study, three earthquake events were selected as the initial medoids: the event at coordinates $(3.967, 95.923)$ for cluster 1, the event at $(-0.386, 98.910)$ for cluster 2, and the event at $(-5.794, 104.037)$ for cluster 3. These initial medoids were then used to compute the distances between every non-medoid object and each medoid using the Euclidean distance as defined in Equation (5). After all distances were calculated, each object was assigned to the cluster of its nearest medoid. This process was repeated until the medoid positions no longer changed (convergence).

After five iterations, the final medoids were obtained and had converged. The final medoids are located at coordinates $(0.115, 97.040)$ for cluster 1, $(-1.8195, 100.389)$ for cluster 2, and $(-3.048, 102.233)$ for cluster 3. The depth and magnitude values (not standardized) for the medoids of each cluster are presented in Table 5.

Table 5. Converged Medoid Data

l-th Cluster Medoid	Depth	Magnitude
1	34.9	5.8
2	68.1	5.2
3	30.8	5.1

The clustering results are summarized in Table 6.

Table 6. Clustering Results for Each Earthquake Event

No.	Longitude	Latitude	Depth	Magnitude	Cluster
1	98.047	2.018	43.1	5.1	3
2	98.755	-1.385	25	5.1	3
3	95.911	4.768	31	5.1	3
4	103.162	-4.772	61.7	5.2	2
5	101.261	-2.825	58.6	5	2
⋮	⋮	⋮	⋮	⋮	
1035	98.754	1.244	92	5.5	2
1036	96.428	1.963	22	5.6	1

No.	Longitude	Latitude	Depth	Magnitude	Cluster
1037	96.033	4.873	5	5.3	3

The K-Medoids clustering produced three clusters, containing 283, 268, and 486 earthquake events, respectively.

Initial Findings

To facilitate an understanding of the characteristics of each cluster, the clusters are described based on the mean values of the two variables presented in Table 7.

Table 7. Cluster Characteristics Based on Variable Means

l-th Cluster	Mean Depth	Mean Magnitude
1	36.8	5.86
2	83.6	5.22
3	31.1	5.15

Based on the mean depth and mean magnitude values for each cluster presented in Table 7, it is evident that each cluster exhibits distinct mainshock earthquake characteristics in terms of depth and magnitude. Cluster 1 is represented by earthquakes with a mean depth of approximately 36.8 km and a magnitude of 5.86, indicating shallow earthquakes with relatively higher magnitudes compared to the other clusters. Cluster 2 is characterized by a medoid depth of about 83.6 km and a magnitude of 5.22, reflecting intermediate-depth earthquakes with moderate magnitudes. Meanwhile, Cluster 3 is marked by earthquakes with a depth of approximately 31.1 km and a magnitude of 5.15, which correspond to shallow earthquakes with relatively lower magnitudes. These differences in medoid values indicate that the clustering process successfully distinguishes mainshock earthquake events based on the combined characteristics of depth and magnitude.

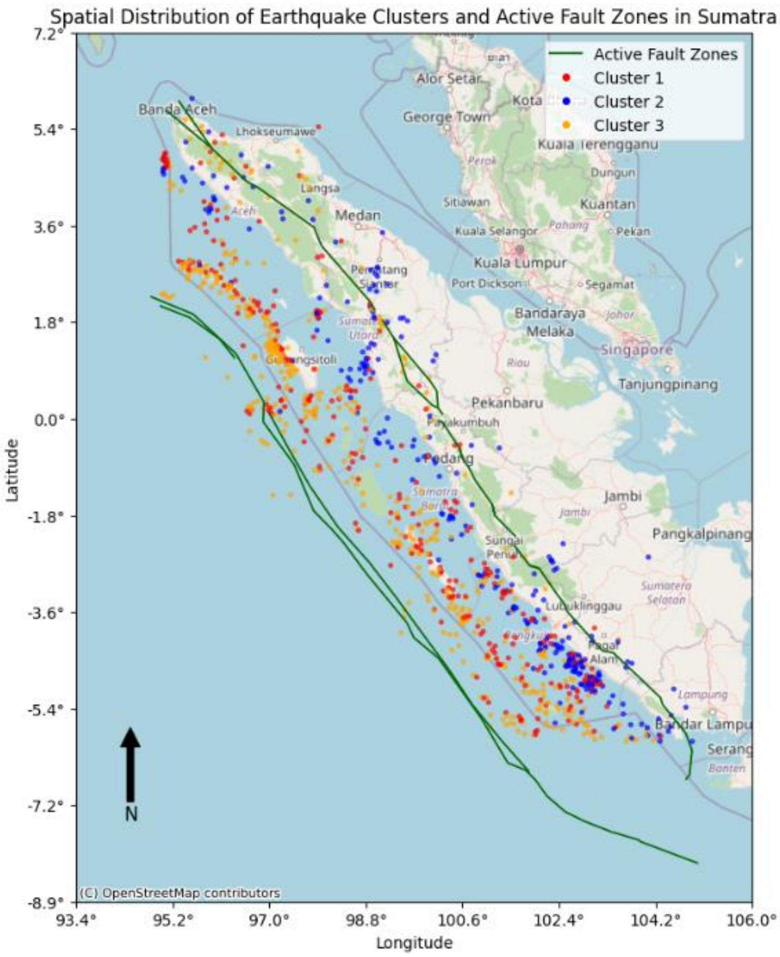


Figure 3. Spatial Distribution of Earthquake Clusters and Active Fault Zones in Sumatra

As illustrated in Figure 3, the spatial distribution of the K-medoids clustering results reveals distinct patterns for each earthquake cluster within the study area, particularly in relation to the active fault zones in Sumatra. Cluster 1 is predominantly concentrated along the main subduction zone offshore the western coast of Sumatra and closely aligns with major active fault segments, indicating a strong association with active plate boundaries and fault-controlled seismicity. Cluster 2 exhibits a relatively more dispersed spatial distribution and is generally located closer to inland areas, where it overlaps with several onshore fault structures, suggesting a relationship with deeper or more complex tectonic features beneath the continental crust. In contrast, Cluster 3 is mainly distributed around the outer subduction zone and follows a pattern parallel to the trench system, with limited overlap with onshore active faults. These differences in spatial distribution, particularly with respect to the active fault zones, indicate variations in the tectonic settings and seismic sources that control earthquake occurrences across the clusters.

Analysis and Discussion

The cluster analysis in this study indicates that grouping mainshock earthquake events based on magnitude and depth can be used as an approach to understand variations in mainshock earthquake characteristics in Sumatra Island. This approach facilitates the interpretation of large mainshock earthquake datasets by simplifying individual events into groups with similar characteristics.

The spatial distribution patterns of the resulting clusters reveal differences in the distribution of mainshock earthquake occurrences across the study area. These differences indicate that mainshock earthquakes in Sumatra Island tend to cluster within specific zones that are closely related to regional tectonic conditions. Accordingly, the clustering results not only describe similarities in earthquake characteristics from a numerical perspective but also demonstrate the relationship between the clustering outcomes and the geological setting of the study area.

From a methodological perspective, the use of the K-Medoids algorithm provides stable clustering results for earthquake data with considerable variability. The selection of medoids as cluster centers reduces the influence of extreme values in the data, particularly in the depth variable. This indicates that K-Medoids is suitable for seismic data, which often exhibit uneven distributions and contain outliers.

Nevertheless, this study has limitations because it only considers two main variables, namely magnitude and depth. Other factors that may influence earthquake characteristics, such as.source mechanisms and temporal relationships between events, are not included in the analysis. Therefore, the results obtained in this study remain descriptive and serve as a basis for further analytical development.

D. Conclusion

This study demonstrates that the K-Medoids method is effective in clustering mainshock earthquake events in Sumatra Island based on magnitude and depth. The evaluation using the Silhouette Coefficient indicates that the optimal number of clusters is three, with a silhouette value of 0.405, reflecting a moderate level of cluster separation. The resulting clustering reveals distinct mainshock earthquake characteristics and different spatial distribution patterns across the study area. This approach can serve as a foundation for further analyses by incorporating additional variables or alternative methods to achieve a more comprehensive understanding of seismic characteristics in Sumatra Island.

References

[1] S. Dey Rahmawati, “Spatial Modeling of Earthquake Risk in Sulawesi and Maluku Based on Geological Factors,” *ESTIMASI: Journal of Statistics and Its Application*, vol. 6, no. 2, pp. 2721–379, 2025, doi: 10.20956/ejsa.v6i2.45160.

[2] K. Choirunnisa, M. N. Fahmi, and A. Realita, “SEISMISITAS PULAU SUMATERA DAN WILAYAH SEKITARNYA AKIBAT AKTIVITAS SEISMO-TEKTONIK ZONA SUBDUKSI,” *Jurnal Inovasi Fisika Indonesia (IFI)*, vol. 14, no. 2, pp. 253–267, 2025.

[3] D. A. N. Sirodj, M. N. Aidi, B. Sartono, U. D. Syafitri, and B. Pranata, “Identification of Earthquake Prone Zones in Sumatra using Density Based Spatial Clustering of Applications with Noise,” *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, vol. 10, no. 2, pp. 916–928, Sep. 2025, doi: 10.18860/cauchy.v10i2.36120.

[4] J. T. Simamora and E. L. Namigo, “Pemetaan Magnitude of Completeness (Mc) untuk Gempa Sumatera,” *Jurnal Fisika Unand*, vol. 5, no. 2, pp. 179–186, Apr. 2016, doi: 10.25077/jfu.5.2.179-186.2016.

[5] A. A. Eluyemi, S. Baruah, and S. Baruah, “Empirical relationships of earthquake magnitude scales and estimation of Guttenberg–Richter parameters in gulf of Guinea region,” *Sci. Afr.*, vol. 6, pp. 1–8, Nov. 2019, doi: 10.1016/j.sciaf.2019.e00161.

[6] Q. Salsabillah and T. Prastowo, “ANALISIS RELASI MOMEN SEISMIK DAN MAGNITUDO MOMEN UNTUK VARIASI KEDALAMAN SUMBER GEMPA TEKTONIK (SHALLOW, INTERMEDIATE, AND DEEP SOURCES),” *Inovasi Fisika Indonesia*, vol. 11, no. 1, pp. 8–16, Feb. 2022, doi: 10.26740/ifi.v11n1.p8-16.

[7] W. Anggraeni, “Analisis Pengelompokan Gempa Bumi di Indonesia Berdasarkan Ruang-Waktu-Kekuatan Kedalaman,” *NUCLEUS*, vol. 4, no. 2, pp. 136–160, Feb. 2024, doi: 10.37010/nuc.v4i2.1480.

[8] U. Rafflesia, D. Rosadi, D. P. Sari, and P. Novianti, “Analysis of Seismic Data in Sumatra using Robust K-Means Clustering,” *Journal of Applied Data Sciences*, vol. 6, no. 1, pp. 391–404, Jan. 2025, doi: 10.47738/jads.v6i1.523.

[9] A. P. Putra, J. Tshivana, and E. Rilvani, “PERBANDINGAN TEORITIS DAN EKSPERIMEN ALGORITMA K-MEANS DAN K-MEDOIDS DALAM KLASTERISASI DATA,” *Kohesi: Jurnal Multidisiplin Saintek*, vol. 10, no. 2, pp. 10–24, 2025, doi: 10.8734/Kohesi.v1i2.365.

[10] M. Orisa and A. Faisol, “Analisis Algoritma Partitioning Around Medoid untuk Penentuan Klasterisasi,” *Jurnal Teknologi Informasi dan Terapan*, vol. 8, no. 2, pp. 86–90, Dec. 2021, doi: 10.25047/jtit.v8i2.258.

[11] A. Mazarei, R. Sousa, J. Mendes-Moreira, S. Molchanov, and H. M. Ferreira, “Online boxplot derived outlier detection,” *Int. J. Data Sci. Anal.*, vol. 19, no. 1, pp. 83–97, Jan. 2025, doi: 10.1007/s41060-024-00559-0.

[12] Effiyaldi *et al.*, “Penerapan Uji Multikolinieritas Dalam Penelitian Manajemen Sumber Daya Manusia,” *Jurnal Ilmiah Manajemen dan Kewirausahaan*, vol. 1, no. 2, pp. 94–102, 2022, [Online]. Available: https://ejournal.unama.ac.id/index.php/jumanage

[13] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[14] M. A. Mercioni and S. Holban, “A Survey of Distance Metrics in Clustering Data Mining Techniques,” in *Proceedings of the 2019 3rd International Conference on Graphics and Signal Processing*, New York, NY, USA: ACM, Jun. 2019, pp. 44–47. doi: 10.1145/3338472.3338490.

[15] D. Marlina, N. Fauzer Putri, A. Fernando, and A. Ramadhan, “Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak,” *Jurnal CoreIT*, vol. 4, no. 2, 2018.

[16] V. Fauziah and Suliadi, “Pengelompokkan Desa di Jawa Barat Berdasarkan Variabel Penyusun Indeks Desa Membangun,” *Bandung Conference Series: Statistics*, vol. 4, no. 1, pp. 221–229, Feb. 2024, doi: 10.29313/bcss.v4i1.11589.

[17] C. Nisa and L. Anggi Kurniawati, “Perbandingan Algoritma K-Means dan K-Medoids dalam Klasterisasi Migrasi Penduduk Kabupaten/Kota di Jawa Tengah,” *Jurnal Riset Statistika*, pp. 131–140, Dec. 2025, doi: 10.29313/jrs.v5i2.8701.

[18] A. F. Nurfauzan and S. Darwis, “Clustering Probabilistic Seismic Hazard Analysis Temporal Epidemic-Type Aftershock Sequence untuk Premi Asuransi,” *Jurnal Riset Statistika*, pp. 41–48, Jul. 2024, doi: 10.29313/jrs.v4i1.3864.