ISSN 2723-6471

Analysis of Seismic Data in Sumatra using Robust K-Means Clustering

Ulfasari Rafflesia[1], Dedi Rosadi[2,*], Devni Prima Sari[3], Pepi Novianti[4]

[1,2,4]*Department of Mathematics, Universitas Gadjah Mada, Yogyakarta, Indonesia*

[1,4]*Department of Mathematics, Universitas Bengkulu, Bengkulu, Indonesia*

[3]*Department of Mathematics, Universitas Negeri Padang, Padang, Indonesia*

(Received: August 18, 2024; Revised: September 5, 2024; Accepted: October 22, 2024; Available online: December 29, 2024)

Abstract

Indonesia is located within the Pacific Ring of Fire and frequently experiences significant seismic activities, rendering the region susceptible to hazards. Specifically, Sumatra is an island in the western part of the country, near the Eurasian and Indo-Australian tectonic plates. Over the past five years, an observable uptick in seismic events has been recorded in Sumatra. This research aimed to cluster the Sumatra region's seismic data using the k-means algorithm and its extensions, including trimmed and robust sparse k-means, to determine the characteristics and patterns of seismic events. The k-means clustering algorithm operates effectively on many data but needs to work better in the presence of outliers. Meanwhile, the data identification reports the presence of outliers in the seismic data. The clustering analysis identified two main clusters, supported by multivariate and spatial outlier detection during preprocessing. The first cluster, encompassing 62% of seismic events, is located offshore near the Mentawai seismic gap, characterized by shallow depths (33–41 km) and magnitudes of 4.5–5.0 Ms. The second cluster, representing 28% of events, includes both mainland and offshore regions, associated with the Sumatran Fault system and slab deformation zones, at moderate depths (54–154 km) with magnitudes of 4.3–4.4 Ms. Rare deep-focus events exceeding depths of 214 km were identified as outliers. Evaluation using Silhouette, Davies-Bouldin, and Dunn indices determined that k=2 was the optimal number of clusters. This study contributes by integrating robust clustering methods to handle outliers, enhancing the reliability of seismic data analysis. This study demonstrates the value of applying trimmed and robust sparse k-means algorithms to improve clustering performance in regions with complex tectonic activity.

Keywords: K-means, Outliers, Robust, Clustering, Earthquake

1. Introduction

Indonesia is among the world's most tectonically active regions due to the location within the Pacific Ring of Fire, an area known for converging three primary tectonic plates, namely the Indo-Australian, Eurasian, and Pacific [1]. Sumatera is positioned at the intersection of the Eurasian and Indo-Australian plates, facilitating the development of geological features such as the Sumatran Fault Zone (SFZ), Subduction Zone, and Mentawai Fault. The SFZ is a tectonic fault that shows a dextral strike-slip motion, extending from the Andaman Sea to the Sunda Strait. Indonesia and a significant portion of Southeast Asia experience regular seismic activity from many sources, rendering the region susceptible to seismic hazards [2]. According to records from 2017 to 2021, seismic data from USGS catalog shows 11,365 earthquakes in Indonesia, with 1,390 events occurring in Sumatra.

Earthquake spatial data contain information about seismic hazard areas within a specified geographical area. This data provides information about places vulnerable to earthquakes and can be used for mitigation processes, to minimize the impact on society and environment. Previous research have focused on establishing the minimal magnitude threshold when conducting seismic hazard assessments [3]. Here, we analyze earthquake data at a place relative to the others using clustering methods, as the extension of the research provided in [4], to determine the characteristics and patterns of data in each cluster formed.

Clustering is actively researched in various fields, including statistics, pattern recognition, and machine learning [5]. The procedure aims to group data based on the concepts of distance and similarity [6]. Furthermore, k-means are the

***Corresponding author: Dedi Rosadi (dedirosadi@ugm.ac.id)**
DOI: https://doi.org/10.47738/jads.v6i1.523

most often used algorithms and play a significant role in data mining [7]. This algorithm can accurately assign labels to the clusters when there are no outliers in the data but does not perform well in the presence of outliers [8]. Based on the given situations, the efficacy of k-means algorithm should be enhanced. Different methods are used to address the negative impact of outliers on the k-means algorithm [9]. Kondo, Barrera, and Zamar [10] proposed robust and sparse k-means to improve Witten and Tibshirani's sparse k-means in 2010 by cutting only a small percentage of the observations at the farthest distance from the cluster's center, using the 1997 Cuesta-Albertos trimmed approach. In [11], a procedure is proposed as a robust and sparse k-means clustering procedure, for high-dimensional data to simultaneously detect groups, outliers, and informative variables to improve performance. Moreover, the robust trimmed k-means method is also introduced to expand the algorithm, which can be used for single or multi-membership data [12]. The k-means algorithm operates iteratively to minimize the within-cluster sum of squares (WCSS) by updating cluster centroids until convergence, which is typically reached when changes in centroids become negligible. Although computationally efficient and suitable for large datasets, its sensitivity to outliers can hinder performance. This research leverages trimmed and robust sparse k-means to overcome these challenges and enhance clustering.

K-means is commonly used in cluster analysis in different areas to identify clusters of earthquake events [4], [13], [14], [15]. The research by [4] used a k-means method to examine the seismic zones in Bengkulu Province. In addition, the algorithm is used to categorize seismic ground-motion records [13]. A modified k-means clustering algorithm partitions global earthquake data into distinct groups within global and local regions [14]. The cluster analysis is used to discover seven seismic zones in Maharashtra State by analyzing earthquake events through a homogenized data [15].

The sensitivity of standard k-means to outliers presents significant challenges when applied to seismic data, which often includes anomalies generated by unexpected geological events or noisy measurements. To address these limitations, this study extends the research in [4] by incorporating robust variants of k-means. Preliminary analysis of Sumatra's seismic data reveals outliers, particularly in Bengkulu and nearby areas, which could skew clustering results if unaddressed. The trimmed and robust sparse k-means methods were selected for their effectiveness in managing outliers and high-dimensional data. Unlike DBSCAN or hierarchical clustering, these methods integrate outlier detection with clustering. Trimmed k-means reduce the impact of extreme outliers by excluding a small portion of anomalous points, while robust, sparse k-means identify clusters, outliers, and key variables, ensuring more reliable and interpretable results.

In this context, this research investigates cluster analysis and related procedures to detect the multivariate and spatial outliers in seismic data, as well as identify patterns used to view tectonically active regions causing earthquakes. The research is structured into sections 1, 2, 3, and 4, providing an introduction overview, method, results and discussion, as well as conclusion, respectively.

2. Literature Review

Clustering has become an essential tool in seismic data analysis to identify spatial and temporal patterns of earthquakes. The k-means method, as one of the most popular clustering algorithms, has been widely used in seismology [4]. The research conducted by [13] demonstrates how k-means clustering can be effectively applied to analyze ground motion recordings due to earthquakes. By using energy distribution in the frequency domain, as well as parameters such as magnitude and propagation distance, this method produces relevant data clustering. Meanwhile, the enhanced k-means clustering technique for global earthquake catalog analysis and earthquake magnitude prediction has also been carried out by [14]. This research highlights the importance of clustering in detecting spatial and temporal patterns in large seismic datasets. Additionally, cluster analysis enabled delineating seven seismic zones in Maharashtra State by examining earthquake events from a standardized dataset [15]. However, k-means is very sensitive to outliers, which often appear in seismic data due to noise or unusual geological events.

Various adaptations of the k-means algorithm have been introduced to mitigate the impact of outliers. The trimmed k-means method mitigates outlier impact by omitting a small subset of data points far from the cluster centroids [12]. The robust sparse k-means technique concurrently discovers clusters, detects outliers, and selects the most pertinent variables [11]. This study’s findings indicate that these resilient methodologies produce clustering outcomes that are more accurate and stable than the conventional k-means algorithm.

 ISSN 2723-6471

However, most studies employed standard k-means without considering outliers, which can substantially influence clustering results, especially in seismic data with considerable variability and noise. The existence of outliers frequently results in inaccurate cluster centroids, misclassification of seismic zones, and poor interpretability of clustering results. Although much research has investigated robust clustering techniques, more comprehensive evaluations of these methods must be conducted, especially in seismic data applications characterized by outliers. The present study addresses deficiencies using the trimmed k-means method and strong sparse k-means, which reduce the influence of outliers and improve the interpretability of clustering results. This study presents a new approach to improve clustering performance in seismic analysis, particularly in areas characterized by complex tectonic activity.

3. Methodology

3.1. Outlier Detection

The initial step in achieving a logical analysis is the detection of outlier observations [16]. An outlier refers to an observation that differs significantly, causing concern of being generated by a different process. Properly investigated outliers provide essential new insight into the results of data analysis. Moreover, the outlier detection for multivariate can be carried out using the Mahalanobis distance [17], which measures the distance of a data point from the mean of a distribution, taking into account the correlation between variables. This distance is evaluated through χ^2 in degrees of freedom equal to the number of variables used. The presence of outliers is reported when the values of the Mahalanobis distances exceed the threshold value of χ^2.

For multivariate data that includes spatial variables, outlier detection can be performed using the mean algorithm. Spatial outlier detection includes the identification of objects dissimilar to spatial neighbors [18]. The Z-value approach measures outliers by calculating the standardized difference between two objects regarding spatial proximity. Spatial outliers are identified when the difference between the threshold value θ and the object in consideration exceeds a given value. In addition, the Z-value is performed through the provided formula.

$$Z_{S(x)} = \left|\frac{S(x) - \mu_s}{\sigma_s}\right| \quad (1)$$

$S(x)$ represents a difference between the non-spatial variable value of object x and the average non-spatial variable value of x's neighboring objects. Meanwhile, μ_s denotes the mean value of $S(x)$ and σ_s represents the standard deviation of $S(x)$ for the entire data.

$S(x)$ represents a difference between the non-spatial variable value of object x and the average non-spatial variable value of x's neighboring objects. Meanwhile, μ_s denotes the mean value of $S(x)$ and σ_s represents the standard deviation of $S(x)$ for the entire data.

Spatial outlier detection based on the mean algorithm involves several steps [19]. Given a set of spatial data $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, the process begins by defining the number k of nearby neighbors, a variable function f, and a threshold $\theta = \chi^2_{s;1-\alpha}$ that has already been set. Each variable f_j is then standardized using the formula $f_j(\boldsymbol{x}_i) \leftarrow \frac{f_j(\boldsymbol{x}_i) - \mu_{f_j}}{\sigma_{f_j}}$, $i = 1,2,\ldots,n$, and j $(1 \leq j \leq q)$. The algorithm proceeds by identifying the k-nearest neighbor set $N, N_k(\boldsymbol{x}_i)$ for each spatial point $\boldsymbol{x}_i$. For each point, a neighborhood function g is determined such that $g_j(\boldsymbol{x}_i)$ represents the mean of the data $\{f_j(\boldsymbol{x}): \boldsymbol{x} \in NN_k(\boldsymbol{x}_i)\}$. To compare the spatial variables, the function $h(\boldsymbol{x}_i) = f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)$ is used. Finally, the algorithm computes $d^2(\boldsymbol{x}_i) = (h(\boldsymbol{x}_i) - \boldsymbol{\mu}_s)^T \sum_s^{-1} (h(\boldsymbol{x}_i) - \boldsymbol{\mu}_s)$ and if $d^2(\boldsymbol{x}_i) \geq \theta$ then $\boldsymbol{x}_i$ is a spatial outlier.

3.2. K-Means Cluster

K-means is an unsupervised learning algorithm frequently used in clustering [20]. This method was initially introduced by MacQueen in 1967 as a partitional hard clustering algorithm [21]. The main objective is to divide objects to be analyzed into distinct and separate clusters [22]. The process involves three main steps. First, the objects are divided into k initial clusters. Next, all objects are listed, and each is assigned to the cluster with the shortest mean distance, using Euclidean distances with either standardized or non-standardized observations. For clusters that gain or lose objects, compute the new centers as follows.

 ISSN 2723-6471

$$C_{kj} = \frac{x_{ikj} + x_{2kj} + \ldots + x_{akj}}{a} \quad (2)$$

where C_{kj} represents the cluster center k of variable j and a is the number of objects in each cluster k. This step should be repeated until no object moved to different cluster.

The Total Within-Cluster Sum of Squares (TWCSS) is a measure used to evaluate the compactness of clusters by calculating the sum of squared distances between each data point and its corresponding cluster center. It represents the variability within each cluster, where a smaller TWCSS value indicates more compact and well-defined clusters. The formula for TWCSS is as follows:

$$\text{TWCSS} = \sum_{i=1}^{N}\sum_{k=1}^{K} I(\mathbf{x}_i \in C_k)\|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2 \quad (3)$$

where $\bar{\boldsymbol{x}}_k$ = cluster center of C_k and $\boldsymbol{I}(\boldsymbol{x}_i \in C_k)$ is 1 if X is true, 0 otherwise. This metric is widely used to assess clustering quality and ensure meaningful group separations.

3.3. Trimmed K-Means Cluster

The trimmed k-means algorithm, introduced by Cuesta-Albertos, Gordaliza, and Matrán [12], is a robust clustering method designed to mitigate the influence of outliers on clustering results. Unlike standard k-means, which can be heavily affected by extreme values, trimmed k-means identifies and removes a proportion of outliers before determining the cluster centers. The process can be summarized in the following steps. The algorithm begins by initializing k cluster centers using a standard initialization method. Then, the squared Euclidean distance to each data point is computed for each cluster center, representing the point's distance from the cluster centers. An outlier trimming process follows, where a trimming proportion α (between 0 and 1) is specified as a parameter. This proportion determines the percentage of points to exclude as outliers, with $\lfloor \alpha N \rfloor$ data points having the largest distances to their closest cluster centers temporarily removed.. If $\alpha = 0.05$, for example, 5% of the total data points are treated as outliers and excluded. Subsequently, the cluster centers are updated by recalculating them based on the remaining data points. The updated cluster center for the k-th cluster ($\boldsymbol{tm}(O)_k$) is computed as:

$$\mathbf{tm}(O)_k = \frac{1}{|C_k \setminus O|}\sum_{i \in C_k \setminus O} \mathbf{x}_i \in \mathbb{R}^p \quad (4)$$

Here, $C_k \setminus O$ represents the instances in cluster k, excluding the trimmed outliers. Therefore, when $O = \emptyset$, then $\boldsymbol{tm}(O)_k = \bar{x}_k$, the method reduces to standard k-means. The algorithm minimizes the within-cluster sum of squares (WSS), defined as:

$$\text{WSS}(C, O) = \sum_{k=1}^{K}\sum_{i \in C_k \setminus O}\|\mathbf{x}_i - \mathbf{tm}(O)_k\|^2 \quad (5)$$

The trimming process ensures that outliers (points with disproportionately high squared distances) do not inflate the WSS. The steps of distance calculation, outlier trimming, cluster center update, and WSS minimization are repeated iteratively until convergence, typically when the changes in cluster centers or WSS fall below a predefined threshold.

3.4. Robust Sparse K-Means Cluster

The robust sparse k-means method was proposed by Kondo, Barrera, and Zamar [10] as a solution for handling big data with outliers. This method combines the trimmed k-means and sparse k-means algorithms to address two major challenges in clustering: the presence of outliers and high-dimensional data.

Sparse k-means, as introduced by Witten and Tibshirani [11], enhance clustering performance in high-dimensional datasets by introducing a sparsity constraint. This mechanism assigns weights (w_j) to each variable, optimizing these weights to minimize the clustering objective function while enforcing sparsity. Variables with higher relevance to the cluster structure are assigned greater weights, while irrelevant variables are down-weighted or excluded. The sparsity constraint improves interpretability by focusing the clustering process on the most meaningful variables.

In robust sparse k-means, the sparse component is integrated to refine the clustering process further, ensuring that only significant variables contribute to the cluster formation. The weights w_j are iteratively updated based on the variability explained by each variable across clusters, as defined by $B_j(C_1, \dots, C_K, O)$.

The robust sparse k-means algorithm can be described as follows: The trimmed k-means is initially applied to the weighted dataset to determine an initial set of cluster centers $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_k)$ and weights (w_j). Next, data points are allocated to clusters based on weighted Euclidean squared distances, as defined by:

$$\min_{C_1,\dots,C_K} \sum_{k=1}^{K} \sum_{i \epsilon C_k} \sum_{j=1}^{p} w_j (x_{i,j} - \mu_{k,j})^2 \quad (6)$$

In the outlier trimming steps, $\alpha * 100\%$ observations with the largest distances to cluster centers are removed and the remaining observations are updated for each cluster. Following this, the sparse weight update step optimizes the weights w_j for the trimmed dataset O and cluster centers, using the formula:

$$\max_{\|w\|_2 \le 1, \|w\|_1 \le l} \sum_{j=1}^{p} w_j B_j(C_1, \dots, C_K, O) \quad (7)$$

where $B_j(C_1, \dots, C_K, O)$, $1 \le j \le p$ measures the variability explained by variable j. The sparsity constraint ensures that irrelevant variables are down-weighted. After this, the cluster centers are recomputed without considering weights, using the partition resulting from trimmed k-means. These steps are repeated iteratively until convergence.

3.5. Clustering Validity

The process of evaluating algorithm results is called the clustering validity process [23]. Validating clustering and determining the correct number of clusters is essential for analysis [24]. Furthermore, clustering validity provides insight into determining the clustering that best fits the data, determining the number of clusters in the data, and providing meaningful results from the clustering. Internal clustering validity measures are commonly used to determine the optimal number of partitions for dividing a dataset [25].

Several cluster validation indices are used to evaluate the quality of clustering results [26]. The silhouette index is the predominant and effective internal validation measure [27]. It is computed as follows.

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}} \quad (8)$$

where $a(i)$ is the average distance between sample i and the other samples within the same cluster, while $b(i)$ reflects the shortest distance between sample i and any other sample in a different cluster [28]. Silhouette index takes values from $\langle -1,1 \rangle$. The maximum index value calculates the best possible clusters within the data.

The Davies–Bouldin (DB) index is a metric used to assess the performance of clustering algorithms. This method measures clustering quality based on inherent data characteristics and factors [29], and the calculation is performed using the following formula.

$$DB = \frac{1}{K} \sum_{k=1}^{K} \max_{i \ne j} \left(\frac{\delta_i + \delta_j}{d_{ij}} \right) \quad (9)$$

The variable d_{ij} represents the distance between the centroids of clusters C_i and C_j. Furthermore, δ_i refers to the standard deviation of the distance of objects in C_i to the centroid of the cluster and a lower DB index value signifies a better clustering solution.

The Dunn index quantifies the ratio of the smallest distance between clusters to the most significant distance within clusters [30] and the index is denoted by

$$\text{Dunn} = \frac{\min_{1 \le i \le j \le K} d(C_i, C_j)}{\max_{1 \le k \le K} \text{diam}(C_k)} \quad (10)$$

where $d(C_i, C_j)$ is the dissimilarity function between two clusters, C_i and C_j, defined as $d(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$.

The diameter of a cluster, denoted as $\text{diam}(C)$, represents the measure of dispersion. The diameter of cluster C can be determined by $\text{diam}(C) = \max_{x,y \in C} d(x, y)$ and the clustering is better when the Dunn index value is higher.

 ISSN 2723-6471

3.6. Data

The seismic data analyzed in this study were obtained from the United States Geological Survey (USGS) earthquake database via the website http://earthquake.usgs.gov. The study focuses on the Sumatra region, geographically defined by latitudes ranging approximately from 0° to -6° and longitudes from 95°E to 106°E. The dataset includes records of earthquakes in this region from January 1st, 2017, to December 31st, 2021. Only earthquakes with a magnitude of 4.0 or greater on the Richter scale (SR) were included to ensure data relevance. The dataset comprises 1390 earthquake events, with variables including latitude, longitude, depth, and magnitude. The USGS earthquake database is a widely recognized and reliable source of global seismic data, ensuring the accuracy and comprehensiveness required for this analysis.

After careful inspection, no missing data were identified in the dataset, as all recorded events were complete. Consequently, no imputation or handling of missing data was required for this analysis. In other scenarios where missing data may occur, established techniques such as multiple imputation, mean substitution, or predictive modelling are typically applied to address this issue and maintain the integrity of the analysis.

The clustering analysis, including k-means and its extensions (robust sparse and trimmed k-means), was conducted using R software (version 4.3.2). The primary libraries utilized for this study were ClusterR for standard k-means and clustering algorithms, RSKC for robust sparse k-means, ktaucenters for trimmed k-means, clusterSim for evaluating cluster validity indices, and factoextra for visualizing and interpreting the clustering results. Additional libraries, such as tidyverse and ggplot2, were employed for data manipulation and visualization. The computations were performed on a laptop with an Intel Core i7-1165G7 processor, 16GB of RAM, and a 512GB SSD running Windows 11. This setup efficiently processed the dataset, which comprised 1390 earthquake events with four variables, completing the clustering analysis and visualizations in under one hour. These resources were adequate for the dataset size; larger datasets may require advanced computational infrastructure or parallel processing techniques.

4. Results and Discussions

4.1. Outlier Detection for Seismic Data

In this study, we employed Mahalanobis distance and spatial outlier detection methods to identify anomalies in seismic data, chosen for their effectiveness in handling seismic data's multivariate and spatially dependent nature. Mahalanobis distance detects multivariate outliers by accounting for correlations between variables such as latitude, longitude, depth, and magnitude, providing a robust measure of deviation from the data's centroid. On the other hand, spatial outlier detection helps identify geographical outliers, such as earthquakes occurring outside expected tectonic plate boundaries, which can indicate unique seismic activity.

These methods outperform traditional techniques like univariate thresholding or clustering approaches by considering multivariate relationships and spatial patterns, enabling a more comprehensive analysis. This approach enhances the ability to capture the underlying patterns in seismic data, improving the reliability of clustering results.

Outliers were detected using Mahalanobis distance with a significance level of $p = 0.95$ and evaluated using using χ^2 at degrees of freedom equal to the number of variables. In this research, four variables were used, since the $\chi^2(k = 4, p = 0.95)$ value was 9.487729 at the $p = 0.95$ level. The Mahalanobis distance and the chi-squared distribution graph quartile were also helpful in detecting outliers. Based on figure 1, there were 94 outliers in Sumatra's earthquake data, which was around 6.76% of the entire data.

 ISSN 2723-6471

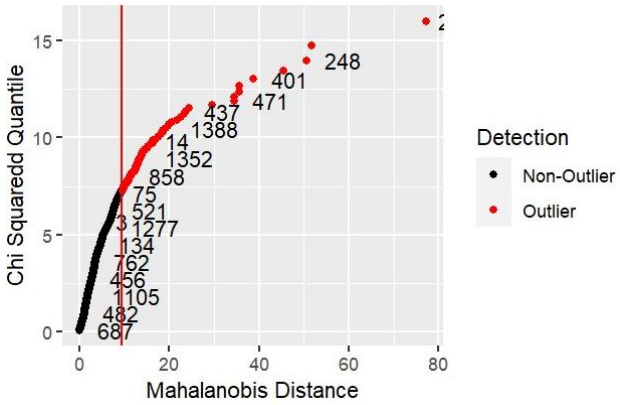

Figure 1. Multivariate Outlier Detection by Mahalanobis Distance

The outlier detection was used by the mean algorithm and the data had spatial and non-spatial dimensions. The spatial dimension of latitude and longitude determined the proximity relationship. Meanwhile, the non-spatial dimension, depth, and magnitude were used to determine the distance function. Based on the detection algorithm, 91 outliers, approximately 6.54% of the data, were classified as spatial outliers. The presence of spatial outliers in data can be observed visually as stars in figure 2.

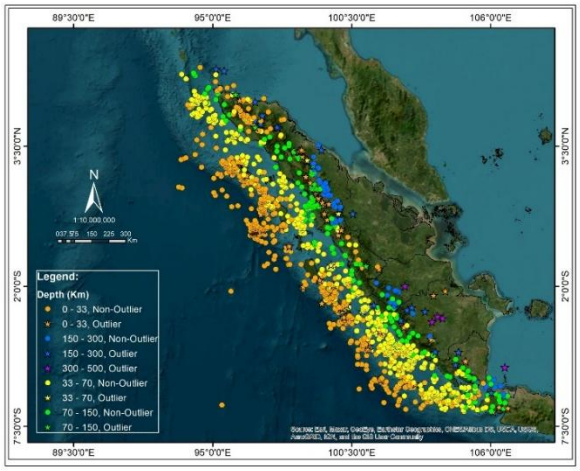

Figure 2. Spatial Outliers Detection for Seismic Data in Sumatra

Based on the observations shown in figure 2, outliers show a range of depths and are distributed over multiple sites. In the context of statistics, outliers refer to instances where the depth or magnitude of an earthquake significantly deviates from the average of neighboring earthquakes. These spatial outliers, characterized by extreme depths or unusual magnitudes, may indicate specific geophysical phenomena such as slab deformation within the subduction zone, interactions between tectonic plates at atypical depths, or rare intraplate seismic activity. Their identification is crucial for understanding distinct tectonic mechanisms and holds significant implications for seismic hazard assessment. These findings can inform targeted studies, improve seismic risk models, and support more effective hazard mitigation strategies in the Sumatran region by highlighting zones of unrecognized seismic activity.

While this study focuses on identifying outliers in seismic data, noise in the dataset was indirectly addressed by using robust methods like Mahalanobis distance and spatial outlier detection. These approaches account for the relationships between variables and spatial patterns, helping minimize random noise's impact. Since the dataset was sourced from a reliable database and showed no clear inconsistencies, noise was not treated as a separate issue. Future studies could explore how noise affects clustering results and consider additional preprocessing steps to improve data quality and reliability for seismic hazard analysis.

 ISSN 2723-6471

Table 1 shows the top ten depth outliers identified using Mahalanobis distance and spatial outliers for seismic data. The 248th earthquake at position (-5.1793, 106.5654) occurred at a depth of 370.43 km with a magnitude of 4.5, identified as an outlier with the deepest earthquake based on the Mahalanobis distance and spatial outlier. The second deepest earthquake occurred at location 493, with a depth of 365.24 and a strength of 4 Ms.

Table 1. Top Ten Depth Outliers Identified Using Mahalanobis Distance and Spatial Outliers

Mahalanobis Distance					Spatial outliers				
Location	**Latitude**	**Longitude**	**Depth**	**Ms**	**Location**	**Latitude**	**Longitude**	**Depth**	**Ms**
248	-5.1793	106.5654	370.43	4.5	248	-5.1793	106.5654	370.43	4.5
493	-3.2574	104.1041	365.24	4	493	-3.2574	104.1041	365.24	4
507	-3.1881	103.8936	345.27	4	507	-3.1881	103.8936	345.27	4
471	-3.3898	103.6835	306.97	4.2	471	-3.3898	103.6835	306.97	4.2
641	-1.9886	102.6106	306.36	4.5	641	-1.9886	102.6106	306.36	4.5
519	-2.9918	103.4033	291.06	4.2	519	-2.9918	103.4033	291.06	4.2
907	0.8464	100.5422	242.96	4.6	907	0.8464	100.5422	242.96	4.6
550	-2.8252	102.887	236.61	4.1	550	-2.8252	102.887	236.61	4.1
528	-2.9353	102.6955	225.65	4.1	528	-2.9353	102.6955	225.65	4.1
1388	6.466	95.4471	224.2	4.3	1388	6.466	95.4471	224.2	4.3

The top ten strength outliers identified using Mahalanobis distance and spatial outliers for seismic data are shown in table 2. The table showed that the 401st location had an earthquake of 6.9 Ms at a depth of 26 km. Meanwhile, the 381st location had an earthquake magnitude of 6.8 Ms at a depth of 22 km.

Table 2. Top Ten Strength Outliers Identified using Mahalanobis Distance and Spatial Outliers

Mahalanobis Distance					Spatial outliers				
Location	**Latitude**	**Longitude**	**Depth**	**Ms**	**Location**	**Latitude**	**Longitude**	**Depth**	**Ms**
401	-4.2069	101.2411	26	6.9	401	-4.2069	101.2411	26	6.9
381	-4.3217	101.1347	22	6.8	381	-4.3217	101.1347	22	6.8
780	0.1364	96.6442	11	6.7	780	0.1364	96.6442	11	6.7
437	-3.7682	101.6228	31	6.4	437	-3.7682	101.6228	31	6.4
694	-1.159	99.6881	43.14	6.3	694	-1.159	99.6881	43.14	6.3
1118	2.3481	96.3575	17	6.3	1118	2.3481	96.3575	17	6.3
187	-5.6856	101.6495	10	6.3	187	-5.6856	101.6495	10	6.3
792	0.1831	96.5601	9	6.1	792	0.1831	96.5601	9	6.1
544	-2.8462	100.0743	20	6	544	-2.8462	100.0743	20	6
566	-2.6706	99.3227	19	6	566	-2.6706	99.3227	19	6

4.2. Optimal Cluster Number

Evaluation of the cluster findings accurately represent the underlying data is a fundamental and critical step of the process. This evaluation uses the Silhouette Index, Davies-Bouldin (DB) Index, and Dunn Index to ascertain the most suitable number of clusters k=2 to k=6, as presented in table 3, table 4, and table 5, respectively.

Silhouette Index, DB, and Dunn are metrics used to evaluate the quality of clusters in clustering analysis. The Silhouette Index measures the level of cohesion and separation between clusters; higher values indicate better-defined clusters. Based on table 3, the highest Silhouette Index value was achieved at k=2 for all clustering methods, indicating that two clusters provide the most optimal separation.

 ISSN 2723-6471

Table 3. Validity Index Values of Silhouette Index

Cluster Number	Robust Sparse K-Means	Trimmed K-Means	K-Means
K=2	0.6429279*	0.5737227*	0.6715969*
K=3	0.5762358	0.5262035	0.5804639
K=4	0.5676858	0.5391605	0.5266975
K=5	0.5939952	0.5596139	0.5457166
K=6	0.5501190	0.5337272	0.6298383

* Indicates the optimal performance for each clustering method based on the Silhouette Index

On the other hand, the DB measures the ratio between intra-cluster distance (within the cluster) and inter-cluster distance; a lower value indicates better cluster quality. Based on table 4, the DB index shows k=2 as optimal for the robust sparse k-means method, while for the trimmed k-means, the best results are obtained at k=5. For the standard k-means method, k=2 provides the smallest DB value, thus supporting the selection of two clusters.

Table 4. Validity Index Values of DB Index

Cluster Number	Robust Sparse K-Means	Trimmed K-Means	K-Means
K=2	0.4655679*	0.5572837	0.4300351*
K=3	0.5298187	0.6079218	0.5311558
K=4	0.5079278	0.5630818	0.6197747
K=5	0.4713902	0.4997942*	0.5749621
K=6	0.5356486	0.5181578	0.4311593

* Indicates the optimal performance for each clustering method based on DB Index

The Dunn Index evaluates the ratio between the minimum inter-cluster distance and the maximum intra-cluster distance; a higher value reflects more separated and compact clusters. From table 5, the Dunn index consistently selects k=2 as the optimal number of clusters for all clustering methods. The combination of results from these three indices indicates that k=2 is the best choice for the number of clusters, providing an optimal balance between cluster compactness and separation.

Table 5. Validity Index Values of the Dunn Index

Cluster Number	Robust Sparse K-Means	Trimmed K-Means	K-Means
K=2	0.0037326*	0.0068071*	0.0079961*
K=3	0.0009617	0.0015983	0.0005866
K=4	0.0017667	0.0021482	0.0020884
K=5	0.0009706	0.0014704	0.0020702
K=6	0.0008253	0.0017214	0.0018086

* Indicates the optimal performance for each clustering method based on Dunn Index

The optimal cluster number for the robust sparse k-means method is k=2 based on all three indices. For the trimmed k-means method, k=2 is optimal according to the Silhouette and Dunn indices, while the DB index suggests k=5. For the k-means method, all indices unanimously support k=2 as the optimal choice. Based on majority voting across these methods and indices, k=2 is the optimal number of clusters. This conclusion ensures robust and interpretable clustering outcomes. The clustering methods are subsequently applied to seismic data using k=2 as the chosen parameter.

It is important to note that determining the optimal number of clusters is inherently influenced by the dataset's characteristics and the specific focus of each validity index. For instance, the Silhouette and Dunn indices prioritize cluster compactness and separation, whereas the DB index evaluates within-cluster tightness relative to inter-cluster distances. These varying criteria can lead to conflicting recommendations, as observed in this study. We adopted a multi-index approach combined with majority voting across three indices and three clustering methods to mitigate

potential bias in cluster selection. While this strategy provides a balanced and robust framework, the chosen value of k ultimately reflects a consensus rather than an absolute measure.

4.3. Clustering Result

The clustering results were visualized using R software, overlaying clusters on a map of Sumatra. The x-axis and y-axis represent longitude and latitude, with colored ellipses indicating clusters identified by the k-means, trimmed and robust sparse k-means algorithms. These ellipses, based on a 95% confidence interval from a t-distribution, outline the spatial distribution of 95% of data points in each cluster, reflecting their spread and density. The varying shapes and sizes of the ellipses highlight differences in the geographical dispersion and concentration of seismic events.

The k-means clustering analysis of seismic events provides a detailed understanding of the seismic activity in Sumatra. Figure 3 shows the outcomes of the clustering analysis conducted on the seismic events in Sumatra using the k-means. The k-means clustering analysis of seismic events in Sumatra reveals two distinct clusters with notable geographical and geological significance. The first cluster, represented by the red ellipse, encompasses 1,214 seismic events with an average magnitude of 4.594 Ms and a depth of 33.458 km. Its center is located offshore near Siberut Island in the Mentawai region (1°14'58.0" S, 99°28'09.1" E), an area influenced by the subduction of the Indo-Australian Plate beneath the Eurasian Plate. This cluster aligns with the tectonic activity of the Mentawai seismic gap, which is known for its shallow earthquakes and tsunami potential.

The second cluster, marked by the blue ellipse, consists of 176 seismic events centered near Tamparungo, Sijunjung, West Sumatra (0°24'41.2" S, 100°52'30.4" E), with an average magnitude of 4.375 Ms and a depth of 154.405 km. The subducting slab's deformation likely influences this deeper seismicity as it bends beneath the Eurasian Plate, a phenomenon typical of subduction zones. Located near the Sumatran Fault system, this cluster suggests interactions between deeper tectonic processes and the fault system, distinguishing it from the first cluster's shallow seismicity linked to the active Mentawai subduction zone.

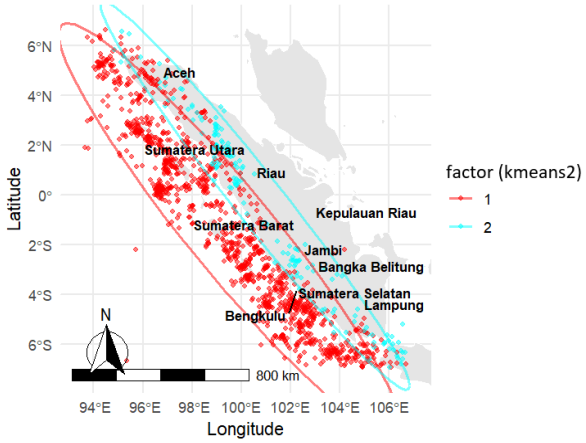

Figure 3. Result of K-Means Cluster with k=2 for Seismic Data

The seismic data analysis using the trimmed k-means algorithm provides further refinement in identifying outliers within the dataset. The trimmed k-means algorithm identifies outliers at 13 distinct places, representing approximately 0.93% of the dataset. The dataset has 13 locations: 248, 318, 471, 493, 507, 519, 528, 550, 641, 907, 1224, 1347, and 1388. The provided data constitutes a subset of outliers discovered by the mean algorithm from the preceding step. Outliers are classified under the second cluster, with a depth beyond 214 km.

Figure 4 shows the cluster results of the Sumatra earthquakes, as obtained from the trimmed k-means clustering analysis. The trimmed k-means algorithm excludes outlier events and identifies two refined clusters that highlight the seismic activity's core patterns. The first cluster (red ellipse) consists of 1,189 seismic events, with an average magnitude of 4.593 Ms and a depth of 33.240 km, centered offshore near Siberut Island in the Mentawai region (1°15'02.6" S, 99°27'53.12 E). This cluster reinforces the prevalence of shallow earthquakes associated with the active subduction zone in this area.

 ISSN 2723-6471

The second cluster (blue ellipse) includes 201 seismic events with an average magnitude of 4.383 Ms and a depth of 153.185 km, centered on the mainland near Tamparungo, Sijunjung, West Sumatra (0°25'17.6" S, 100°52'26.8" E). These deeper seismic events are likely linked to the bending and deformation of the subducting Indo-Australian Plate as it moves into the mantle, resulting in intermediate-depth earthquakes influenced by the plate's shape and its interaction with the overriding Eurasian Plate. The cluster's proximity to the Sumatran Fault system may indicate combined effects of intraplate deformation and stress concentrations near the fault. Additionally, 13 outlier events with depths exceeding 214 km were identified, representing rare high-depth seismic occurrences that deviate from the main patterns. Unlike the first cluster, which is associated with shallow subduction in the Mentawai seismic gap, this cluster reveals more complex tectonic activity.

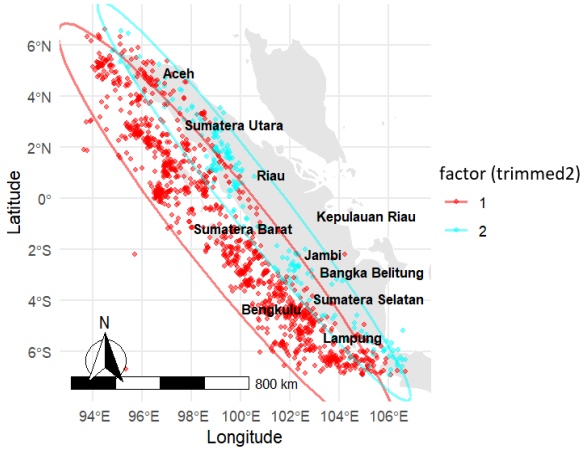

Figure 4. Result of Trimmed K-Means Cluster with k=2 for Seismic Data

The robust sparse k-means algorithm provides a comprehensive seismic data analysis, offering valuable insights into earthquake clustering in Sumatra. Figure 5 shows the cluster results obtained through the robust sparse k-means algorithm in Sumatra. The first cluster is denoted by the red elliptical region with a total of 1189 seismic events. The center of the first cluster lies at a latitude of 1°28'54.9" S and a longitude of 99°50'20.1 E, within proximity of Siberut and the Mentawai Islands. The mean magnitude of the earthquakes in the initial cluster was 4.906 Ms, with a depth of 41.454 km.

The oval section with a blue color corresponds to the second cluster and contains 201 earthquakes. In addition, the center of the cluster is located at 0°53'45.0" S and 99°30'26.0" E. This geographical area is situated within the ocean between the islands of Siberut, Mentawai, West Sumatra, and Sumatra. The mean magnitude of the earthquakes in the second cluster was 4.317 Ms, with a mean depth of 54.122 km.

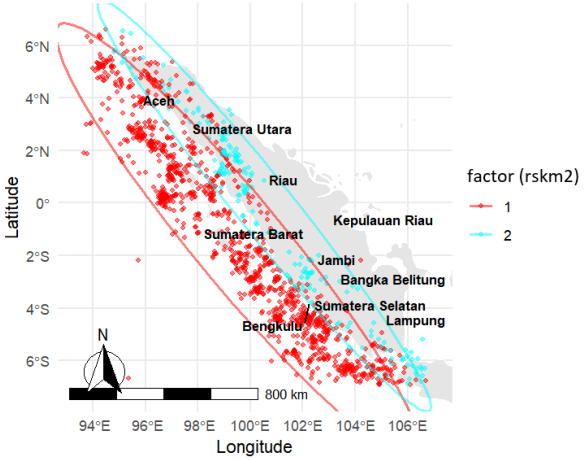

Figure 5. Result of Robust Sparse K-Means Cluster with k=2 for Seismic Data

The moderate-depth earthquakes in the second cluster may be linked to subduction processes where the slab begins to deform and interact with mantle materials. Its proximity to the Sumatran Fault system suggests possible stress redistribution between the slab and nearby tectonic structures, contrasting with the shallow seismicity of the first cluster near the Mentawai subduction zone. The robust sparse k-means algorithm identified 13 data points as outliers, all exceeding depths of 200 km, which likely represent rare deep-focus events associated with slab deformation or localized mantle dynamics. These outliers within the second cluster highlight its association with greater depths compared to the cluster center.

Although the visual representations of clusters across the methods appear similar due to the two-dimensional projection, the differences in cluster membership, centroids, and seismic characteristics highlight the unique contributions of each clustering technique.

5. Conclusion

In conclusion, this study applied the k-means algorithm and its extensions (trimmed k-means and robust sparse k-means algorithms) to cluster seismic data from Sumatra Island, Indonesia. The clustering analysis identified two clusters based on spatial and geophysical characteristics, supported by multivariate and spatial outlier detection during data preprocessing. Evaluation using the Silhouette index, DB index, and Dunn index determined that k=2 was the optimal number of clusters.

The first cluster predominantly comprises seismic events in oceanic areas near the Mentawai seismic gap, influenced by the subduction of the Indo-Australian Plate beneath the Eurasian Plate. These events, characterized by shallow depths (33–41 km) and magnitudes of 4.5 to 5 Ms, primarily occur offshore but also include some extending to coastal land regions. The second cluster spans mainland and offshore regions, with seismic events concentrated near the Sumatran Fault system and slab deformation zones at moderate depths (54–154 km). These earthquakes, magnitudes of 4.3 to 4.4 Ms, reflect complex tectonic processes, including intraplate deformation and slab-mantle interactions. A few high-depth outliers exceeding 214 km were identified, representing rare deep-focus events.

This study did not validate the clustering results using external data, such as independent seismic datasets, known seismic hazard zones, or previous studies. Although the clustering results align with the general tectonic framework of Sumatra, future research could validate the clusters using external data. This would help confirm consistency with established seismic hazard patterns or geological studies, improving the accuracy and reliability of the findings.

While this study focused on clustering based on spatial and geophysical variables, future research could explore additional factors, such as earthquake frequency over time or historical seismic data. Incorporating these variables may reveal temporal patterns and trends, offering deeper insights into the seismic activity in Sumatra.

6. Declarations

6.1. Author Contributions

Conceptualization: U.R., D.R., D.P.S., and P.N.; Methodology: U.R., and D.R.; Software: U.R., and P.N.; Validation: U.R., D.R., and D.P.S.; Formal Analysis: U.R., D.R., and D.P.S.; Investigation: U.R., and D.R.; Resources: U.R., and P.N.; Data Curation: U.R., and P.N.; Writing – Original Draft Preparation: U.R., D.R., and D.P.S.; Writing – Review and Editing: U.R., D.R., D.P.S., and P.N.; Visualization: U.R. and P.N.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received financial support for the research from the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan/LPDP).

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Irsyam, P. R. Cummins, M. Asrurifak, L. Faizal, D. H. Natawidjaja, S. Widiyantoro, I. Meilano, W. Triyoso, A. Rudiyanto, S. Hidayati, and M. Ridwan, "Development of the 2017 national seismic hazard maps of Indonesia," *Earthquake Spectra*, vol. 36, no. 1_suppl, pp. 112–136, 2020, doi: 10.1177/8755293020951206.

[2] S. J. Hutchings and W. D. Mooney, "The Seismicity of Indonesia and Tectonic Implications," *Geochemistry, Geophys. Geosystems*, vol. 22, no. 9, pp. 1–42, 2021, doi: 10.1029/2021GC009812.

[3] A. A. S. Putra, B. A. D. Nugraha, C. N. T. Puspito, and D. D. P. Sahara, "Preliminary result: Source parameters for small-moderate earthquakes in Aceh segment, Sumatran fault zone (Northern Sumatra)," in 18th Annual Meeting of the Asia Oceania Geosciences Society, Aug. 2022, no. AOGS 2021, pp. 224–226, doi: 10.1142/9789811260100_0076.

[4] P. Novianti, D. Setyorini, and U. Rafflesia, "K-means cluster analysis in earthquake epicenter clustering," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 2, pp. 81–89, Jul. 2017, doi: 10.26555/ijain.v3i2.100.

[5] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng. Appl. Artif. Intell.*, vol. 110, no. December 2021, pp. 104743, 2022, doi: 10.1016/j.engappai.2022.104743.

[6] S. Askari, "Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Syst. Appl.*, vol. 165, no. August 2020, pp. 113856, 2021, doi: 10.1016/j.eswa.2020.113856.

[7] R. Mussabayev, N. Mladenovic, B. Jarboui, and R. Mussabayev, "How to Use K-means for Big Data Clustering?," *Pattern Recognit.*, vol. 137, no. Mei 2023, pp. 109269, 2023, doi: 10.1016/j.patcog.2022.109269.

[8] N. H. M. M. Shrifan, M. F. Akbar, and N. A. M. Isa, "An adaptive outlier removal aided k-means clustering algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6365–6376, 2022, doi: 10.1016/j.jksuci.2021.07.003.

[9] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, vol. 622, no. April 2023, pp. 178–210, 2023, doi: 10.1016/j.ins.2022.11.139.

[10] Y. Kondo, M. Salibian-Barrera, and R. Zamar, "RSKC: An R package for a robust and sparse k-means clustering algorithm," *J. Stat. Softw.*, vol. 72, no. 5, pp. 1-26, 2016, doi: 10.18637/jss.v072.i05.

[11] Š. Brodinová, P. Filzmoser, T. Ortner, C. Breiteneder, and M. Rohm, "Robust and sparse k-means clustering for high-dimensional data," *Adv. Data Anal. Classif.*, vol. 13, no. 4, pp. 905–932, 2019, doi: 10.1007/s11634-019-00356-9.

[12] O. Dorabiala, J. N. Kutz, and A. Y. Aravkin, "Robust trimmed k-means," *Pattern Recognit. Lett.*, vol. 161, no. September 2022, pp. 9–16, Sep. 2022, doi: 10.1016/j.patrec.2022.07.007.

[13] Y. Ding, Y. Peng, and J. Li, "Cluster Analysis of Earthquake Ground-Motion Records and Characteristic Period of Seismic Response Spectrum," *J. Earthq. Eng.*, vol. 24, no. 6, pp. 1012–1033, 2020, doi: 10.1080/13632469.2018.1453420.

[14] R. Yuan, "An improved K-means clustering algorithm for global earthquake catalogs and earthquake magnitude prediction," *J. Seismol.*, vol. 25, no. 3, pp. 1005–1020, 2021, doi: 10.1007/s10950-021-09999-8.

[15] S. Dhole and S. Bakre, "An updated homogeneous earthquake catalogue and earthquake recurrence parameters of Maharashtra state, an Indian stable continental region," *J. Earth Syst. Sci.*, vol. 133, no. 13, pp. 1-25, 2024, doi: 10.1007/s12040-023-02220-z.

[16] A. Smiti, "A critical overview of outlier detection methods," *Comput. Sci. Rev.*, vol. 38, no. November 2020, pp. 100306, 2020, doi: 10.1016/j.cosrev.2020.100306.

[17] H. Ghorbani, "Mahalanobis Distance and Its Application for detecting multivariate outliers," *Facta Univ.*, vol. 34, no. 3, pp.

583–595, 2019.

[18] S. Shekhar, C. T. Lu, and P. Zhang, "A unified approach to detecting spatial outliers," GeoInformatica, vol. 7, no. 2, pp. 139–166, 2003.

[19] C.-T. Lu, D. Chen, and Y. Kou, "Multivariate spatial outlier detection," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 4, pp. 801–811, Dec. 2004, doi: 10.1142/S021821300400182X.

[20] S. S. Yu, S. W. Chu, C. M. Wang, Y. K. Chan, and T. C. Chang, "Two improved k-means algorithms," *Appl. Soft Comput. J.*, vol. 68, no. July 2018, pp. 747–755, 2018, doi: 10.1016/j.asoc.2017.08.032.

[21] R. Garcia-Dias, C. A. Prieto, J. S. Almeida, and I. Ordovás-Pascual, "Machine learning in APOGEE," *Astron. Astrophys.*, vol. 612, no. A98, pp. 1-56, Apr. 2018, doi: 10.1051/0004-6361/201732134.

[22] E. U. Oti, M. O. Olusola, F. C. Eze, and S. U. Enogwe, "Comprehensive Review of K-Means Clustering Algorithms," *Int. J. Adv. Sci. Res. Eng.*, vol. 07, no. 08, pp. 64–69, 2021, doi: 10.31695/ijasre.2021.34050.

[23] Q. Xu, Q. Zhang, J. Liu, and B. Luo, "Efficient synthetical clustering validity indexes for hierarchical clustering," *Expert Syst. Appl.*, vol. 151, no. August 2020, pp. 1-13, Aug. 2020, doi: 10.1016/j.eswa.2020.113367.

[24] F. Ros, R. Riad, and S. Guillaume, "PDBI: A partitioning Davies-Bouldin index for clustering evaluation," *Neurocomputing*, vol. 528, no. April 2023, pp. 178–199, 2023, doi: 10.1016/j.neucom.2023.01.043.

[25] M. Gagolewski, M. Bartoszuk, and A. Cena, "Are cluster validity measures (in) valid?," *Inf. Sci. (Ny).*, vol. 581, no. December 2021, pp. 620–636, Dec. 2021, doi: 10.1016/j.ins.2021.10.004.

[26] N. Wiroonsri, "Clustering performance analysis using a new correlation-based cluster validity index," *Pattern Recognit.*, vol. 145, no. January 2024, pp.109910, Jan. 2024, doi: 10.1016/j.patcog.2023.109910.

[27] X. Wang and Y. Xu, "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index," in *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 569, no. 5, pp. 1-12, Jul. 2019, IOP Publishing.

[28] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[29] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[30] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.