Performance Evaluation of K-Means Clustering Algorithm Using Some Robust Distances: a Case Study on Seismic Data in Sumatra

Ulfasari Rafflesia[1] | *Universitas Gadjah Mada, Yogyakarta, Indonesia*
Dedi Rosadi[2] | *Universitas Gadjah Mada, Yogyakarta, Indonesia*
Devni Prima Sari[3] | *Universitas Negeri Padang, Padang, Indonesia*
Adhitya Ronnie Effendie[4] | *Universitas Gadjah Mada, Yogyakarta, Indonesia*

Received 28.10.2024 (revision received 13.2.2025), **Accepted** (reviewed) 15.4.2025, **Published** 12.12.2025

Abstract

Clustering is an unsupervised learning technique that categorizes data into groups based on inherent patterns and similarities, with K-means being one of the most common methods. K-means clustering is particularly susceptible to outliers because of its dependence on non-robust distances (such as the most used Euclidean distance). To address this issue, robust distance metrics such as a new Standardized Euclidean Robust distance and Mahalanobis Robust distance have been discussed in this paper, which will reduce the influence of outliers and, at the same time, improve clustering accuracy empirically. The main objective of this study is to investigate the impact of applying robust distance metrics in the K-means clustering and to identify the most suitable distance metric for seismic data containing outliers. The findings indicate that robust distance measures outperform the non-robust distances in accuracy, yielding superior outcomes for minimum-valued indices such as Davies-Bouldin, Xie-Beni, and Ball-Hall indices, as well as maximum-valued indices such as Calinski-Harabasz and Dunn indices.

Keywords	DOI	JEL code
Clustering, k-means, outliers, robust distance	*https://doi.org/10.54694/stat.2024.72*	*C38, C55*

[1] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia. E-mail: ulfasarirafflesia@mail.ugm.ac.id, phone: (+62)816353340. ORCID: <https://orcid.org/0000-0002-7739-3952>.

[2] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia. Corresponding author: e-mail: dedirosadi@ugm.ac.id, dedirosadistat91@gmail.com, phone: (+62)81328471898. ORCID: <https://orcid.org/0000-0003-2689-253X>.

[3] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Padang, Padang 25132, Indonesia. E-mail: devniprimasari@fmipa.unp.ac.id, phone: (+62)85868648474. ORCID: <https://orcid.org/0000-0003-0382-8973>.

[4] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia. E-mail: adhityaronnie@ugm.ac.id, phone: (+62)8157965042. ORCID: <https://orcid.org/0000-0002-0840-6318>.

INTRODUCTION

Clustering is an important technique in machine learning and data mining, allowing the partitioning of large datasets into meaningful groups based on intrinsic similarities. Among the many clustering algorithms available, K-means is essential in data mining and is the most straightforward and commonly utilized method (Mussabayev et al., 2023).

K-means clustering algorithm can be used with various distances metric, where the most popular one is using Euclidean distance, see e.g. Patel and Mehta (2012), Kapil and Chawla (2016). Other popular distance including Mahalanobis distance, such as in Nelson (2012) which integrated the Mahalanobis distance into the traditional K-means clustering algorithm. The experiment's results show the advantages of using the Mahalanobis distance in clustering. Brown (2022) discuss improvements to the traditional K-means clustering algorithm by incorporating Mahalanobis distance. They provide experimental results that demonstrate the effectiveness of this method compared to standard Euclidean distance k-means. Ghazal et al. (2021) explores the effects of various distance metrics on the performance of the K-means clustering algorithm. Each metric is evaluated for its performance in various datasets, providing insight into how different distance measurements handle high-dimensional or complex data.The study shows that distance metrics (Euclidean, Manhattan, Mahalanobis, and Chebyshev distances) have a significant influence on the performance of K-means clustering, with the Mahalanobis-Euclidean distance yielding the best accuracy compared to other metrics.

The K-means clustering algorithm is prone to several challenges in practical applications. This algorithm is particularly sensitive to outliers (e.g. Zhang et al., 2021). Outliers can be broadly defined as data instances that significantly diverge from established norms within a data set or from anticipated behavioral concepts (Smiti, 2020). More generally, outliers are data points that exhibit a significant deviation from most of the dataset and can distort clustering results. A robust clustering method is required to mitigate the impact of outlier data (Huber and Ronchetti, 2011).

In this study, we extend the result from Ghazal et al. (2021) by studying K-means clustering using several new robust distance metrics, namely Standardized Euclidean Robust and Mahalanobis Robust distance. This research aims to investigate the impact of these robust distance metrics on clustering accuracy and evaluate their effectiveness in handling seismic data outliers from Sumatra Island. Specifically, this study addresses the following research objectives:

1. To assess whether robust distance metrics improve clustering accuracy in K-means clustering when applied to seismic data containing outliers.
2. To identify the most effective distance metric for clustering seismic data by comparing clustering outcomes based on multiple internal evaluation indices.

To achieve these objectives, we formulate the following hypotheses:

1. K-means clustering with robust distance metrics (Standardized Euclidean Robust or Mahalanobis Robust) will yield higher clustering accuracy than traditional Euclidean distance.
2. The clustering results using robust distance metrics will produce more compact and well-separated clusters, as indicated by superior values in internal evaluation indices such as Davies-Bouldin, Xie-Beni, Calinski-Harabasz, Dunn, and Ball-Hall indices.

The rest of the paper is structured as follows. Section 1 introduces the background of the problem. Section 2 discusses the methods used in the study. Finally, Sections 3 and 4 present the results and discussion, followed by the conclusion at the end of the paper.

1 LITERATURE SURVEY

Clustering is an important technique in data mining and machine learning, especially for grouping data that does not have explicit labels. The K-means algorithm has become one of the most commonly used methods for clustering due to its ease of implementation and computational efficiency (Lantz, 2019).

Several related and important results can be summarized as follows. Maulik and Bandyopadhyay (2002) comprehensively evaluate several clustering algorithms and the effectiveness of different clustering validity indices. This paper compares the performance of K-Means, agglomerative hierarchical clustering (single linkage), and a simulated annealing-based clustering technique in terms of their ability to partition data accurately. The study uses Euclidean distance for K-means and simulated annealing algorithms to compute the squared distance between points and cluster centroids. Additionally, the paper explores the role of various validity indices, including the Davies-Bouldin index, Dunn's index, Calinski-Harabasz index, and a newly developed index I, which measures the quality of clustering results across diverse datasets.

Many studies have used distance metrics like Euclidean and Manhattan. Patel and Mehta (2012) enhanced the K-means algorithm by evaluating Euclidean, Manhattan, and Minkowski distances. Euclidean distance proved more effective for clusters of similar size, whereas Manhattan distance was better for noisy, high-dimensional data. Minkowski provided adaptability but required precise parameter adjustment. Kapil and Chawla (2016) examined Euclidean and Manhattan distances, demonstrating that Euclidean distance was more efficient, requiring fewer repetitions, resulting in lower the within-cluster sum of squared errors (WSS), and exhibiting a faster runtime, proving more effective for their datasets.

Raeisi and Sesay (2022) proposed a new distance metric to improve K-means clustering, particularly for different cluster sizes. Based on Canberra distance, this metric generates larger clusters for centroids that are farther distant from the origin and smaller clusters for those that are nearer. Compared to conventional metrics such as Euclidean and Manhattan, simulations indicate that it generates more precise clusters, which is especially beneficial in domains like autonomous cars. The research determines that the novel metric improves K-means efficacy in irregular data distributions.

Brown (2022) presented a non-trivial integration of K-means clustering and Mahalanobis distance. The method enhances the K-means clustering process by replacing the standard Euclidean distance metric with the Mahalanobis distance. This modification allows the algorithm to account for the covariance among variables in the dataset. The evaluation results indicate that the Mahalanobis distance-based K-means clustering algorithm exhibits superior accuracy compared to comparable studies.

K-means has a major weakness: its sensitivity to outliers because it uses the average to determine the centroid and its dependence on non robust distances (such as the mostly used Euclidean distance). This weakness is making it vulnerable to distortions caused by outliers or noise in the data. Outliers can significantly affect clustering results and reduce accuracy (Zhao, Ying and Karypis, 2002).

Various approaches have been developed to address this issue. Hawkins (1994) introduces the Feasible Solution Algorithm (FSA) for computing the Minimum Covariance Determinant (MCD) estimator, which is crucial for robust multivariate data analysis. Hardin and Rocke (2004) extend the MCD estimator to the multiple cluster setting for robust outlier detection. This paper develops a method that effectively identifies outliers across multiple clusters using Mahalanobis-type distances based on robust estimates of location and scatter derived from the MCD. Hubert and Debruyne (2010) highlight the MCD estimator‘s efficacy in identifying multivariate outliers. They examine its extensions and applications in topics such as robust principal component analysis and regression. The MCD‘s effectiveness in various fields, such as finance and image analysis, makes it an essential tool in robust statistical methods. Hubert et al. (2018) comprehensively review the MCD estimator, a highly robust method for estimating multivariate location and scatter while effectively managing outliers. The paper discusses the fast algorithm available for computing the MCD, along with its various applications and extensions in fields such as classification, clustering, and other areas of applied and methodological multivariate statistics, demonstrating its versatility in computational statistics.

2 METHODS

2.1 Distance measure

The distance measure must be determined before clustering. The distance measure reflects the degree of separation between data points and should align with the characteristics used to differentiate clusters within the dataset (Cao et al., 2012; Huang, 2008)

2.1.1 Euclidean distance

Euclidean distance is the primary measure used in cluster analysis to quantify the distance between data objects and the centroids of their respective clusters (Johnson and Wichern, 2002). Euclidean distance, D_e, between two data points x_i and x_j is described as:

$$D_e\left(x_i, x_j\right) = \left(\sum_{k=1}^{d} \left|x_{ik} - x_{jk}\right|^2\right)^{\frac{1}{2}}, \tag{1}$$

where x_{ik} and x_{jk} each represent the k-th dimension of x_i and x_j and d is the number of dimensions (Xu and Wunsch, 2005).

2.1.2 Minskowski distance

The Minkowski distance is a generalization of several well-known distance metrics, such as Euclidean and Manhattan distances. It is a parameterized distance metric in a normed vector space determined by a parameter p. Depending on the p-value, the Minkowski distance can represent different distance measures. The Minkowski distance between two data points x_i and x_j in d-dimensional space is given by:

$$D_{Mink}\left(x_i, x_j\right) = \left(\sum_{k=1}^{d} \left|x_{ik} - x_{jk}\right|^p\right)^{\frac{1}{p}}, \tag{2}$$

where x_i and x_j are the two data points, d is the number of dimensions, p is the parameter that controls the type of distance metric: when $p = 1$, this becomes the Manhattan distance, when $p = 2$, it becomes the Euclidean distance and when $p \rightarrow \infty$, it becomes the Chebyshev distance.

2.1.3 Manhattan distance

The Manhattan distance between two data points is defined as the sum of the absolute differences of their coordinates. This distance is also known as a city block, square, taxi, or $L1$ distance. Mathematically, the distance between x_i and x_j is defined as:

$$D_{Mn}\left(x_i, x_j\right) = \sum_{k=1}^{d} \left|x_{ik} - x_{jk}\right|. \tag{3}$$

2.1.4 Mahalanobis distance

The Mahalanobis distance is the alternative distance applied considering the correlation effect between variables. The Mahalanobis distance is defined as the distance between two points that involves the covariance or correlation between variables. The Mahalanobis distance between two objects is expressed in the form of vectors and matrices:

$$D_{Mh}\left(x_i, x_j\right) = \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \boldsymbol{S}^{-1} \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right), \tag{4}$$

with $\boldsymbol{S}$ being the sample covariance matrix.

2.1.5 Canberra distance

Canberra distance measure quantifies the amount of absolute fractional difference between the variables of a pair of data points. This measurement is defined as follows:

$$D_{Can}\left(x_i,x_j\right)=\sum_{k=1}^{d}\frac{\left|x_{ik}-x_{jk}\right|}{\left|x_{ik}\right|+\left|x_{jk}\right|}, \tag{5}$$

where x_{ik} and x_{jk} are components of vectors x_i and x_j, and d is the dimension of the vector. This measurement is sensitive to small changes when both coordinates approach zero.

2.1.6 Chebyshev distance

The Chebyshev distance between two points and in a multi-dimensional space is mathematically defined as:

$$D_{Cb}\left(x_i,x_j\right)=\underset{1\le k\le d}{max}\left(\left|x_{ik}-x_{jk}\right|\right), \tag{6}$$

where: x_i and x_j are two points in a d-dimensional space, x_{ik} and x_{jk} are the k-th feature (coordinate) of the points x_i and x_j, respectively and the maximum is taken over all dimensions k from 1 to d.

2.1.7 Standardized Euclidean distance

The standardized Euclidean distance is defined as the Euclidean distance between data points divided by their standard deviation. The standardized Euclidean distance between x_i and x_j is mathematically represented as:

$$D_{se}\left(x_i,x_j\right)=\left(\sum_{k=1}^{d}\left|\frac{x_{ik}-x_{jk}}{s_k}\right|^2\right)^{\frac{1}{2}}, \tag{7}$$

with s_k being the standard deviation of dimension k (Xu and Tian, 2015).

2.2 Robust distance measure

Outliers can substantially affect distance measures and other statistical computations in data analysis, particularly multivariate data. Robust methods such as the Minimum Covariance Determinant (MCD) minimize the effect of outliers and improve distance calculation. This section discusses the Minimum Covariance Determinant (MCD) method, both with the Standard Euclidean Robust Distance and the Mahalanobis Robust Distance based on MCD.

2.2.1 Minimum Covariance Determinant (MCD)

The Minimum Covariance Determinant (MCD) method is a robust statistical technique used to estimate the covariance matrix and mean of multivariate data (Hubert and Debruyne, 2010). In 1984, Peter Rousseeuw introduced it as a robust alternative to traditional covariance estimation techniques subject to outliers. Within the framework of robust distance, the MCD technique provides a robust covariance estimate, subsequently employed to compute the distance between two data points.

This method aims to find a subset of size h from the entire set of observations, where the covariance matrix of the subset has the smallest determinant among all possible combinations. The subset S_h is assumed to contain data that does not include outliers.

$$\hat{\mu}_{MCD} = \frac{1}{h}\sum_{i=1}^{h} x_i. \tag{8}$$

This is the estimated mean for the subset S_h with the smallest determinant using the MCD.

$$\hat{\Sigma}_{MCD} = \frac{1}{h-1}\sum_{i=1}^{h}\left(x_i - \hat{\mu}_{MCD}\right)\left(x_i - \hat{\mu}_{MCD}\right)^T. \tag{9}$$

This is the covariance estimate for the subset S_h with the smallest determinant, calculated using the MCD.

2.2.2 Standardized Euclidean Distance Robust (using MCD)

The robust standardized Euclidean distance between x_i and x_j in a dataset with d dimensions is given by:

$$D_{robust\_se}\left(x_i, x_j\right) = \left(\sum_{k=1}^{d}\left(\frac{\left|x_{ik} - x_{jk}\right|}{\sqrt{\hat{\sigma}_{k,rob}^2}}\right)^2\right)^{\frac{1}{2}}, \tag{10}$$

where: x_i and x_j are the two data points, each with d features, x_{ik} and x_{jk} are the values of the k-th feature (dimension) for data points x_i and x_j, respectively and $\hat{\sigma}_{k,rob}^2$ is the robust estimate of the variance for the k-th feature, computed using the Minimum Covariance Determinant (MCD) method.

2.2.3 Mahalanobis Robust Distance (using MCD)

The robust Mahalanobis distance between two points x_i and x_j is given by:

$$D_{M,rob}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)^T \Sigma_{rob}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right), \tag{11}$$

where: x_i and x_j are points in d-dimensional space and is the robust covariance matrix estimated from the data.

2.3 Cluster validity

The process of assessing the outcomes of clustering algorithms is referred to as cluster validity (Halkidi, 2001). Internal criteria are crucial in clustering as they evaluate clustering quality independent of external information or labels.

2.3.1 The Ball-Hall index

The Ball-Hall index measures points‘ mean dispersion (variance) within clusters. It is defined as the mean of the sum of squared distances between each point and the centroid of its cluster:

$$BH_{index} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in C_k} x_i - c_k^{\ 2}, \tag{12}$$

where: K is the number of clusters, n_k represents the number of points within the k-th cluster, C_k is the set of points belonging to the k-th cluster, x_i represents the i-th data point within the k-th cluster, c_k is the centroid of the k-th cluster and $||x_i - c_k||^2$ is the squared Euclidean distance between the point x_i and the centroid c_k**.**

2.3.2 The Calinski-Harabasz index

The Calinski-Harabasz index $S(k)$ for a given number of clusters k is calculated as:

$$S(k)=\frac{B(k)/(k-1)}{W(k)/(n-k)}, \tag{13}$$

were: $W(k)$ is the within-cluster sum of squares (unexplained variance) and defined as follows:

$$W(k)=\sum_{i=1}^{k}\sum_{x\in C_i} x-c_i^2, \tag{14}$$

where: k is the number of clusters, C_i is the t-th cluster, x represents data point in cluster C_i, c_i is the centroid of cluster C_i and $||x - c_i||^2$ is the squared Euclidean distance between a point x and the centroid c_i. $B(k)$ is the between-cluster sum of squares (explained variance) and formed as:

$$B(k)=\sum_{i=1}^{k} n_i \|c_i-c\|^2, \tag{15}$$

where: n_i is the number of data point in cluster C_i, c_i is the centroid of cluster C_i, c is the global centroid (centroid of all data points) and $||c_i - c||^2$ is the squared Euclidean distance between the centroid of cluster C_i and the global centroid c.

2.3.3 The Davies-Bouldin index

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 (Davies and Bouldin, 1979), and the calculation is performed using the following formula:

$$DB=\frac{1}{K}\sum_{k=1}^{K}\max_{i\neq j}\left(\frac{\delta_i+\delta_j}{d_{ij}}\right). \tag{16}$$

The variable $d_{ij} = ||c_i - c||^2$, represents the distance between the centroids of clusters C_i and C_j, δ_i refers to the standard deviation of the distance of objects in C_i to the centroid of the cluster and δ_j refers to the standard deviation of the distance of objects in C_j to the centroid of the cluster. A lower DB index value signifies a better clustering solution.

2.3.4 The Det-Ratio index

Let W denote the pooled within-cluster covariance matrix, computed as a weighted sum of individual cluster covariance matrices, and let T represent the total covariance matrix of the entire dataset. The Det-Ratio index calculates the ratio of the determinants of W and T. The index compares the entire dataset‘s overall dispersion (spread) with the spread within individual clusters (Milligan and Cooper, 1985). The Det-Ratio index is computed as:

$$Det-Ratio=\frac{det(W)}{det(T)}, \tag{17}$$

where: $det(W)$ is the determinant of the pooled within-cluster covariance matrix and $det(T)$ is the determinant of the total covariance matrix of the dataset. Formula for $det(W)$:

$$det(W) = \sum_{i=1}^{k} (n_i - 1)\Sigma_i \, , \tag{18}$$

where: k is the number of clusters, n_i represents the number of points in i-th cluster, Σ_i is the covariance matrix of i-th cluster, W is the pooled within-cluster covariance matrix, and $det(W)$ is the determinant of this pooled matrix. Formula for $det(T)$:

$$det(T) = (n-1)\Sigma \, , \tag{19}$$

where: n is the total number of data points in the dataset, Σ is the covariance matrix of the entire dataset, T is the total covariance matrix, and $det(T)$ is the determinant of this total matrix.

2.3.5 The Dunn index

The Dunn index quantifies the ratio of the smallest distance between clusters to the most significant distance within clusters (Dunn, 1973) and the index is denoted by:

$$Dunn = \frac{\underset{1 \le i \le j \le K}{min} \, d\left(C_i, C_j\right)}{\underset{1 \le k \le K}{max} \, diam\left(C_k\right)} \, , \tag{20}$$

where:

- K is the number of clusters.
- $d(C_i, C_j)$ is the distance between clusters C_i and C_j, defined as the minimum distance between any point in C_i and any point in C_j: $d(C_i, C_j) = \underset{x \epsilon C_i, y \epsilon C_i}{min} \|x - y\|$.
  This is the minimum inter-cluster distance.
- $diam(C)$ is s the diameter of cluster C_k defined as the maximum distance between any two points within the same cluster: $diam(C) = \underset{x,\, y \epsilon C_k}{max} \|x - y\|$.
  This is the maximum intra-cluster distance, representing the spread or dispersion within the cluster.

The clustering is better when the Dunn index value is higher.

2.3.6 The Silhouette index

The Silhouette index evaluates how well each data point x_i fits into its assigned cluster by comparing its cohesion within the cluster and its separation from the nearest neighboring cluster. The index is computed as:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(b(x_i) - a(x_i)\right)}{max\left\{a(x_i), b(x_i)\right\}} \, , \tag{21}$$

where:

- n is the total number of data points.
- Cohesion: $a(x_i)$ is the average distance between x_i and all other points in the same cluster C_i:

$$a(x_i) = \frac{1}{|C_i| - 1} \sum_{x_j \epsilon C_i,\, j \neq i} d(x_i, x_j) \, . \tag{22}$$

Separation: $b(x_i)$ is the minimum average distance between x_i and all points in the nearest cluster C_k:

$$b(x_i) = \min_{C_k \neq C_i} \left(\frac{1}{|C_k|} \sum_{x_j \in C_k} d(x_i, x_j) \right), \tag{23}$$

$max\{a(x_i), b(x_i)\}$ ensures that the larger of the two values $a(x_i)$ or $b(x_i)$ is used to normalize the result.

Silhouette index ranges from –1 to 1, the maximum index value calculates the best possible clusters within the data.

2.3.7 Xie-Beni index (XB)

The Xie-Beni index is an index of fuzzy clustering, but it is also applicable to crisp clustering. It is defined as the quotient between the mean quadratic error and the minimum of the minimal squared distances between the points in the clusters (Xie and Beni, 1991). For crisp clustering, this is equivalent to the within-cluster sum of squares (WGSS) divided by the total number of points N. It represents the average squared distance between the data points and the centroid (barycenter) of their assigned cluster.

The inter-cluster distance $\delta_1(C_k, C_k,)$ is calculated as the minimum distance between the points of two clusters C_k and C_k, based on their centroids. The formula is:

$$XB_{index} = \frac{1}{N} \frac{WGSS}{min_{k<l} \delta_1 (C_k, C_l)^2}, \tag{24}$$

where:

- N is the total number of data points.
- $WGSS$ is the within-cluster sum of squares, which is the sum of the squared distances between points in each cluster and their corresponding cluster centroids.

$$WGSS = \sum_{i=1}^{K} \sum_{x \in c_k} \|x - c_k\|^2, \tag{25}$$

where:

- c_k is the k-th cluster, x is a point in cluster c_k, and $||x - c_k||^2$ is the squared Euclidean distance between point x and the centroid c_k,
- $\delta_1(C_k, C_k,)$ is the minimum inter-cluster distance, which is the minimum distance between the centroids of clusters c_k and c_l.

$$\delta_1(c_k, c_l,) = ||c_k - c_l||, \tag{26}$$

where: c_k and c_l are the centroids of clusters C_k and C_l, respectively.

2.4 Spatial outlier detection

A mean-based approach can be used to detect outliers in multivariate data with spatial variables (Lu et al., 2004). This technique discovers spatial outliers by recognizing items that significantly deviate from their neighboring spatial locations (Shukla et al., 2021).

Steps for Spatial Outlier Detection Based on the Mean Algorithm.

- Input Data and Setup: define the spatial data $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$, a fixed number k of nearby neighbors, a function *f*, and a threshold $\theta = \chi^2_{s;1-\alpha}$ that has already been set, where s represents the degrees of freedom (df) in the chi-square distribution.
- Standardization: standardize each variable values of each spatial point.
- Nearest neighbors: identify the *k*-nearest neighbors set *N*, $N_k(\boldsymbol{x}_i)$ for each spatial point $\boldsymbol{x}_i$.
- Neighborhood function: calculate neighborhood function g for each spatial point $\boldsymbol{x}_i$ such that $g_j(\boldsymbol{x}_i)$ = mean of the data $\left\{f_j\left(\boldsymbol{x}\right):\boldsymbol{x}\in NN_k\left(\boldsymbol{x}_i\right)\right\}$ and the function used for comparing $h\left(\boldsymbol{x}_i\right) = f\left(\boldsymbol{x}_i\right) - g\left(\boldsymbol{x}_i\right)$.
- Outlier determination: compute the squared difference $d^2\left(\boldsymbol{x}_i\right)$, $d^2\left(\boldsymbol{x}_i\right) = \left(h\left(\boldsymbol{x}_i\right) - \boldsymbol{\mu}_s\right)^T \sum_s^{-1}\left(h\left(\boldsymbol{x}_i\right) - \boldsymbol{\mu}_s\right)$, and compare it to the threshold θ. If $d^2\left(\boldsymbol{x}_i\right) \geq \theta$ is considered a spatial outlier.

2.5 Modified K-means algorithm

1. Initialization:
    - Centroid initialization: choose an initial set of centroids.
    - Prepare for Distance Metric Selection: Euclidean, Standardized Euclidean, Maximum, Manhattan, Canberra, Minskowski, Mahalanobis and standardized Euclidean Robust and Mahalanobis Robust (using Minimum Covariance Determinant).
2. Distance calculation (modification from procedure 2):
    - Euclidean Distance: for standard k-means.
    - Manhattan, Canberra, etc., can be implemented depending on the structure of your data.
    - Mahalanobis Distance: involves calculating the covariance matrix, or for robust k-means, the MCD estimator.
    - Robust Distance Metrics like MCD, ensure that the robust covariance matrix (or other robust statistics) is used in the distance calculations.
3. Cluster assignment: based on the distance calculations, assign each point to the closest centroid.
4. Centroid update: recalculate the centroids based on the current cluster assignments.
5. Convergence check:

    Repeat steps 2 to 4 until the centroids stabilize (no change between iterations) or the algorithm reaches the predefined number of iterations.

    Modified Step 6: Determining Optimal k , the number of clusters, (from Procedure 3):

    – Loop over k values: run the clustering algorithm for various values of k.

    – For each value of k, calculate the following internal indices:

    o Davies-Bouldin Index: lower values are better.

    o Silhouette Score: higher values indicate better clustering.

    o Calinski-Harabaz, Dunn, Xie-Beni, Ball-Hall, Determinant-Ratio.

    – After computing these indices for different values of k , select the value of k that provides the most favorable results across the indices.

3 RESULTS

In this study, we applied K-means clustering to Sumatra earthquake data using various distance metrics.

3.1 Data description

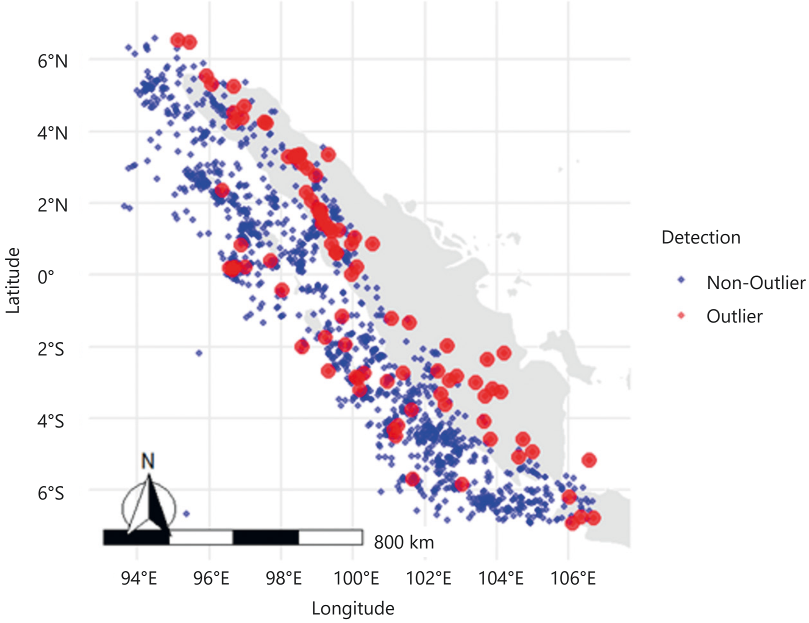
We collected earthquake data for Sumatra from January 1, 2017 to December 31, 2021.[5] We apply the preprocessing by only includes 1 390 recorded earthquakes with a magnitude greater than 4.0 on the

[5] From the website: <*http://earthquake.usgs.gov*>.

Richter scale and it contains variables such as latitude, longitude, depth, and magnitude. The K-means algorithm with several alternative robust distance measures implemented in R software version 4.3.2.

Before applying clustering, it is essential to detect whether spatial outliers exists in the data, as these outliers could significantly affect the accuracy of the clustering results. Outliers, especially in spatial data, represent unusual events or anomalies that deviate from general patterns and can provide essential insights. Therefore, using the mean algorithm (see Section 2.4), we detected outliers by considering the earthquake data‘s spatial (latitude, longitude) and non-spatial (depth, magnitude) dimensions. Ninety-one points, or 6.54% of the dataset, were identified as spatial outliers. Figure 1 illustrates the spatial location of outliers detected in the Sumatra earthquake data.

Figure 1 Spatial outliers detection in earthquake data



Source: Own research

3.2 Clustering accuracy

After detecting outliers, we applied the K-means clustering algorithm to the dataset using several distance metrics. The distance metrics evaluated include Euclidean, standardized Euclidean, robust standardized Euclidean, maximum, Manhattan, Canberra, Minkowski, Mahalanobis, and robust Mahalanobis.

Before performing clustering with the K-means algorithm, it is essential to determine the optimal number of clusters (k). The determination of the optimal k is carried out by comparing various internal indices, such as the Ball-Hall Index, Calinski-Harabasz Index (CH), Determinant Ratio (Det-Ratio), Davies-Bouldin Index (DB), Silhouette, Dunn, and Xie-Beni. The final k value is chosen based on the majority vote across these indices, selecting the k that is most frequently indicated as optimal by the majority of the indices for each distance metric.

Clustering accuracy was then evaluated by assessing the quality of the clusters generated by the algorithm. This was done by comparing the above internal indices, which provide different perspectives on the clustering structure. Table 1 presents the results of calculating these internal indices for each distance metric with k values varying from 2 to 6.

Table 1 Share of positive answers to job search questions and item-response probabilities

Distances	*k*	Ball-Hall	CH	DB	Det-Ratio	Dunn	Silhouette	Xie Beni
Euclidean	*k* = 2	1.273672	686.887	0.997194	2.71645	0.005846	0.400210	665.959
	k = 3	0.936657	1 289.818	0.732494	8.02887	0.005964	0.448821	548.981
	k = 4	0.748148	1 121.184	0.917919	12.50254	0.002291	0.368531	3 526.045
	k = 5	0.646471	1 036.303	0.908646	15.67819	0.001618	0.342675	6 291.175
	k = 6	0.621035	1 142.601	0.830398	25.76036	0.001520	0.360806	5 955.146
Standardized Euclidean	*k* = 2	1.273672	686.887	0.997194	2.71645	0.005846	0.400210	665.9598
	k = 3	0.936823	1 289.825	0.732525	8.03081	0.005964	0.448170	548.979
	k = 4	0.880946	1 013.581	0.870280	10.48739	0.003484	0.375172	1 483.779
	k = 5	0.639051	1 041.950	0.879309	16.00830	0.004813	0.329858	753.451
	k = 6	0.594372	1 037.654	0.921148	21.89493	0.000901	0.321425	1 489.52
Standardized Euclidean robust	*k* = 2	1.397652	797.894	0.760961	3.34890	0.003209	0.496180	2 048.134
	k = 3	0.975659	1 304.917	0.712285	8.32247	0.016157	0.457789	71.438
	k = 4	0.801777	1 076.272	0.964211	12.94352	0.002619	0.360394	2 352.817
	k = 5	0.682322	976.5731	0.975242	16.30381	0.001531	0.324290	6 003.763
	k = 6	0.630923	1 045.075	0.923768	23.43386	0.002007	0.336698	2 794.037
Maximum	*k* = 2	1.841707	519.720	1.382222	2.96414	0.001457	0.310419	6 341.465
	k = 3	0.944584	1 285.413	0.733669	8.01995	0.011600	0.443202	135.182
	k = 4	0.933743	1 031.092	0.811268	10.76380	0.007502	0.408614	348.757
	k = 5	0.634958	1 017.798	0.873809	15.39508	0.002024	0.335769	4 197.289
	k = 6	0.556546	951.830	0.877183	19.53233	0.001866	0.319942	4 541.747
Manhattan	*k* = 2	1.284324	684.613	1.003226	2.72209	0.006441	0.398207	549.133
	k = 3	0.956509	1 280.447	0.734501	7.98718	0.005851	0.446828	551.587
	k = 4	0.760856	1 106.612	0.911945	12.42045	0.001191	0.368243	11 167.37
	k = 5	0.698208	1 108.580	0.897459	17.28403	0.003368	0.345153	1 127.893
	k = 6	0.591964	1 022.215	0.921821	21.65069	0.002572	0.322703	2 272.086
Canberra	*k* = 2	1.673566	266.011	1.830434	1.40051	0.001084	0.163452	24 360.80
	k = 3	1.020687	807.530	0.906802	4.57418	0.000870	0.317712	22 330.27
	k = 4	0.930501	636.572	1.207194	5.48530	0.000400	0.251529	11 5581.2
	k = 5	0.849292	531.193	1.353911	6.22975	0.001291	0.212897	10 390.41
	k = 6	0.663275	513.754	1.157161	8.26076	0.000984	0.204716	15 881.63
Minkowski	*k* = 2	1.273672	686.887	0.997194	2.71645	0.005846	0.400210	665.959
	k = 3	0.936823	1 289.825	0.732525	8.03081	0.005964	0.448170	548.979
	k = 4	0.880946	1 013.581	0.870280	10.48739	0.003484	0.375172	1 483.779
	k = 5	0.639051	1 041.950	0.879309	16.00830	0.004813	0.329858	753.451
	k = 6	0.594372	1 037.654	0.921148	21.89493	0.000901	0.321425	18 489.52
Mahalanobis	*k* = 2	1.843878	520.086	1.380569	2.96581	0.001337	0.310156	7 526.662
	k = 3	0.950738	1 309.881	0.714538	8.30701	0.015120	0.462011	86.154
	k = 4	0.753744	1 103.287	0.943658	12.40621	0.002268	0.363824	3 566.359
	k = 5	0.645699	1 070.142	0.858568	17.07804	0.001902	0.335614	4 199.140
	k = 6	0.649216	1 115.020	0.832299	24.79341	0.002714	0.355516	1 733.396
Robust Mahalanobis	*k* = 2	1.407947	793.578	0.765406	3.35663	0.007641	0.494316	361.282
	k = 3	0.977242	1 301.184	0.714473	8.34197	0.016157	0.456395	71.572
	k = 4	0.87639	1 181.072	0.714400	12.33241	0.005851	0.441421	441.457
	k = 5	0.710398	914.041	1.088614	17.30273	0.001054	0.304277	13 278.71
	k = 6	0.710398	914.041	1.088614	17.30273	0.001054	0.304277	13 278.71

Source: Own research

3.3 Evaluation of internal indices

Based on the internal index evaluation results in Table 1, this can be clarified as follows:

1. For Euclidean distance, at $k = 2$, several indices show pretty good results, such as Ball-Hall with a value of 1.273672 (the smaller, the better) and Davies-Bouldin (DB) with a value of 0.9971942 (the smaller, the better), indicating that the clusters have good compactness. The Calinski-Harabasz (CH) index is at a value of 686.8874 (the more significant, the better), which indicates a reasonable cluster separation, although it can still be improved. However, the Silhouette value of 0.4002105 and the Xie-Beni value of 665.9598 indicate that the clusters still need to be fully optimal for stronger separation and compactness.

   At $k = 3$, the index results show a significant improvement. The Ball-Hall value decreased to 0.936657, which means the clusters became more compact. The CH value skyrocketed to 1 289.818, indicating a much better cluster separation than $k = 2$. Additionally, the DB value decreased to 0.7324939, which indicates better-separated clusters, and the Silhouette increased to 0.4488218, meaning the cluster quality improved. The Xie-Beni index value is also lower at $k = 3$, at 548.9815, indicating that the clusters have better compactness and separation.

   From most indices, $k = 3$ is the optimal choice for Euclidean Distance. A higher CH index, a lower DB index, and an increased Silhouette value indicates that the clusters‘ separation and compactness are better at $k = 3$ than $k = 2$, resulting in more effective and accurate clustering.
2. For the Standardized Euclidean, Standardized Euclidean Robust, Maximum, Manhattan, and Minkowski distances, the majority of indices, such as Ball-Hall, Calinski-Harabasz (CH), Davies-Bouldin (DB), Silhouette, and Xie-Beni, consistently show that $k = 3$ is the optimal solution. Higher CH indices, lower DB indices, and increased Silhouette values across all these metrics indicate better separation and compactness of clusters at $k = 3$, similar to the results found with the Euclidean distance. However, Canberra and Robust Mahalanobis are slightly different, with less optimal results than other distance metrics in choosing $k = 3$, indicating variations that need to be considered in cluster separation.
3. For the Canberra and Mahalanobis distances, the Ball-Hall, Det-ratio, and Silhouette index values yield an optimal k value of $k = 3$. Meanwhile, for the CH, Dunn, and Xie Beni index values, the optimal k value is $k = 2$. Therefore, the optimal value for both distances can be $k = 2$ or $k = 3$.
4. Meanwhile, for the Manhattan distance, based on the majority value of the indices, namely the BH index, Det-Ratio, Dunn, and Xie Benie, the optimal k is $k = 2$.

3.4 Clustering visualization

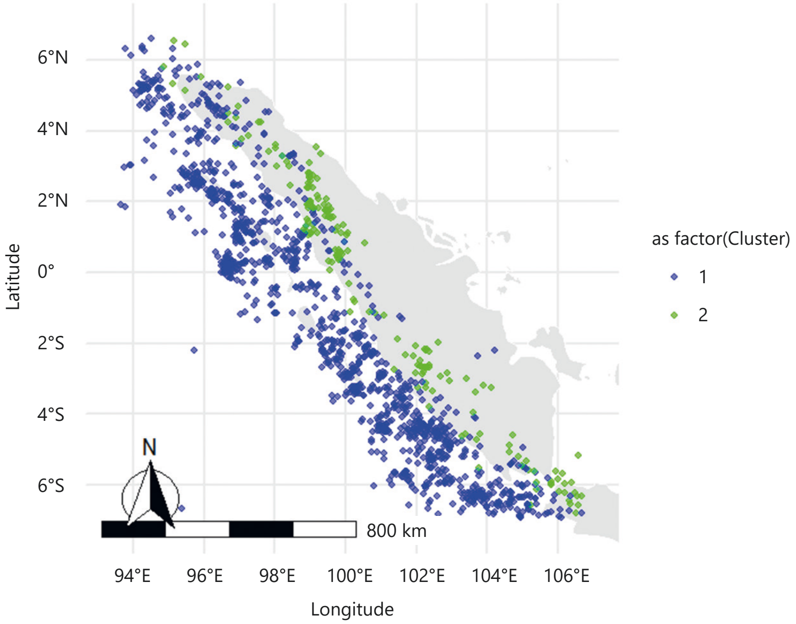
After determining that $k = 2$ and $k = 3$ are the optimal values based on evaluating internal indices using various distance methods, the next step is to display the clustering results for both k values. The clustering visualization results will be used to analyze the characteristics of the formed clusters and evaluate how variations in the number of clusters affect data separation in multidimensional space.

Figure 2 shows the spatial visualization of the clustering results formed for $k = 2$ using the K-means algorithm with the robust, standardized Euclidean distance measure. This visualization identifies distinct spatial patterns between the two clusters based on earthquake magnitude, depth, and spatial distribution variations.

Figure 3 shows the clustering results were shown geographically when $k = 3$ and the K-means algorithm with the robust Mahalanobis distance measure was used. The clustering results indicate more precise distinctions among the three clusters, which will be analyzed further regarding earthquake depth, magnitude, and dispersion patterns. The spatial separation of these clusters enhances the understanding of the geological processes that may be driving the clustering.

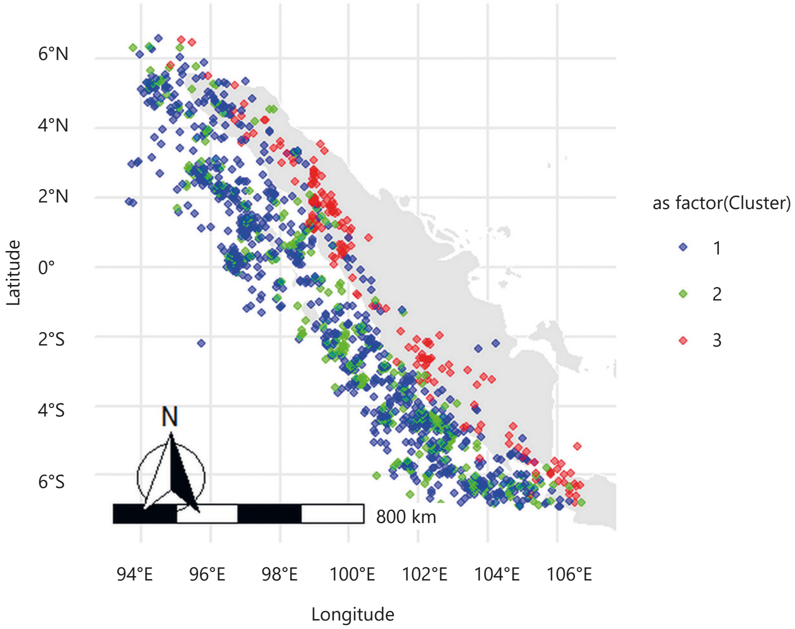
This visualization analysis provides a foundation for further interpretation and discussion of the earthquakes' characteristics in the following chapter.

Figure 2 Result of K-means clustering for *k* = 2 using Standardized Euclidean Robust



Source: Own research

Figure 3 Result of K-means clustering for *k* = 3 using Mahalanobis Robust Distance



Source: Own research

4 DISCUSSION

The analysis considered the earthquake's depth, magnitude, distribution, and implications.

4.1 Analysis of cluster characteristics (*k* = 2)

Earthquakes in Cluster 1 have a much shallower depth, averaging around 33 km, compared to Cluster 2, which has an average depth of 153.7 km. This difference may reflect that the earthquakes in Cluster 1 are closer to the subduction zone or shallow faults around Sumatra, while Cluster 2 is associated with deeper earthquakes, possibly from different tectonic activity or from deeper crustal depths.

The earthquakes in Cluster 1 have a slightly higher average magnitude (4.59 Ms) compared to Cluster 2 (4.40 Ms). Although the difference is small, it suggests that the earthquakes in Cluster 1 may be more destructive due to occurring at shallower depths and with greater magnitude.

Cluster 1 represents more than 85% of all recorded earthquakes, which means that most seismic activity in Sumatra occurs at shallow depths. This is important because shallow earthquakes generally have greater potential to cause structural damage compared to deep earthquakes, whose energy may be absorbed before reaching the surface. Cluster 2, which has deeper earthquakes with lower magnitudes, is likely associated with deeper seismic events in the Earth's crust that are not as hazardous to surface infrastructure.

Based on the analysis results above, using the K-means clustering method based on standardized Euclidean robust distance, two main earthquake clusters in Sumatra have been successfully identified. The first cluster shows shallow earthquakes with slightly larger magnitudes, while the second cluster is more oriented towards deeper earthquakes with lower magnitudes. This has important implications for earthquake risk mitigation, where special attention needs to be given to earthquakes in Cluster 1 as they have the potential to be more destructive.

4.2 Analysis of cluster characteristics (*k* = 3)

The earthquakes in cluster 1 are shallow (about 33 km) with small to moderate magnitudes (around 4.40), indicative of tectonic earthquakes occurring at shallow depths. This cluster comprises 62% of all known earthquakes. Cluster 1 is characterized by shallow earthquakes of low to moderate magnitudes, indicating surface tectonic activity that produces lower-energy seismic events.

Conversely, the earthquakes in cluster 2 indicate a depth similar to that of cluster 1, measuring 34.61 km; nevertheless, this cluster possesses a greater magnitude, around 5.08, compared to the others. Cluster 2 encompasses earthquakes of greater magnitude, representing around 24% of all recorded earthquakes, indicative of greater subduction activity or significant faults that generate more powerful earthquakes, which may be a problem in disaster risk evaluation.

Cluster 3 comprises deep earthquakes (about 155 km) with small to moderate magnitudes. Earthquakes at this depth typically occur in deep subduction zones or plates descending significantly (Benioff zone), resulting in less surface impact. Despite occurring at significant depths, the earthquake's magnitude was minor.

Based on the analysis results above, using the robust Mahalanobis-based K-means clustering method, three main earthquake clusters in Sumatra have been successfully identified. Clusters 1 and 2 are likely related to the earthquakes occurring near the subduction zone, which is the meeting point between the Indo-Australian and Eurasian plates, as well as the Sumatra Fault. These shallow earthquakes may occur in the transitional area between the two tectonic features but the dominance of shallow earthquakes and larger magnitudes in Cluster 2 is likely due to subduction activity. Cluster 3, which consists of deep earthquakes, is most likely directly related to the deep subduction zone, particularly in the Benioff zone. The significantly greater depth of these earthquakes compared to other clusters indicates that they occur due to the movement of plates descending to great depths.

CONCLUSION

Based on the results obtained in this study, robust distance measures, such as Standardized Euclidean Robust and Mahalanobis Robust, effectively handled outliers in seismic data from Sumatra, resulting in more accurate and well-separated clusters. The findings confirmed that these robust approaches consistently outperformed traditional Euclidean distance in clustering accuracy, as demonstrated by various internal evaluation indices.

This study uses K-means cluster analysis with robust metrics like Standardized Euclidean Robust Distance and Mahalanobis Distance to understand where and how earthquakes occur in Sumatra. These metrics effectively accommodate data variation, particularly when handling frequent outliers in seismic data. More representative clusters unaffected by outliers were identified by adopting a robust approach in K-means. This contributes to a better understanding of the tectonic dynamics of the Sumatra region and provides practical implications for disaster risk mitigation. This approach ensures that outliers do not distort the results, leading to a more stable and reliable cluster understanding of seismic patterns.

This study successfully addressed the research objectives, demonstrating that robust distance metrics enhance clustering accuracy and effectively mitigate the impact of outliers. For future research, we recommend further exploration of other robust distance metrics and adaptive clustering algorithms to improve accuracy and robustness in datasets prone to outliers. Extending this approach to high-dimensional and noisy data domains, such as environmental monitoring and market segmentation, could validate its broader applicability. Additionally, dynamic clustering techniques that adaptively update cluster centers based on evolving seismic patterns are suggested to enhance the model‘s adaptability.

References

BROWN, P. O., CHIANG, M. C., GUO, S., JIN, Y., LEUNG, C. K., MURRAY, E. L., PAZDOR, A. G. M., CUZZOCREA, A. (2022). Mahalanobis Distance Based K-Means Clustering [online]. In: *Data & Knowledge Engineering*, 146: 256–262. <https://doi.org/10.1007/978-3-031-12670-3_23>.

CAO, F., LIANG, J., LI, D., BAI, L., DANG, C. (2012). A dissimilarity measure for the k-Modes clustering algorithm [online]. *Knowledge-Based Systems*, 26: 120–127. <https://doi.org/10.1016/j.knosys.2011.07.011>.

DAVIES, D. L., BOULDIN, D. W. (1979). A Cluster Separation Measure [online]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI*, 1(2): 224–227. <https://doi.org/10.1109/TPAMI.1979.4766909>.

DUNN, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters [online]. *Journal of Cybernetics*, 3(3): 32–57. <https://doi.org/10.1080/01969727308546046>.

GHAZAL, T. M., HUSSAIN, M. Z., SAID, R. A., NADEEM, A., HASAN, M. K., AHMAD, M., KHAN, M. A., NASEEM, M. T. (2021). Performances of k-means clustering algorithm with different distance metrics [online]. *Intelligent Automation and Soft Computing*, 30(2): 735–742. <https://doi.org/10.32604/iasc.2021.019067>.

HALKIDI, M. (2001). *On Clustering Validation Techniques* [online]. Springer, pp. 107–145. <http://link.springer.com/article/10.1023/A:1012801612483>.

HAWKINS, D. M. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data [online]. *Computational Statistics and Data Analysis*, 17(2): 197–210. <https://doi.org/10.1016/0167-9473(92)00071-X>.

HUANG, A. (2008). Similarity measures for text document clustering. *New Zealand Computer Science Research Student Conference, NZCSRSC 2008 Proceedings*, April, pp. 49–56.

HUBER, P. J., RONCHETTI, E. M. (2011). *Robust statistics* [online]. John Wiley & Sons. <https://books.google.co.id/books?hl=en&lr=&id=j1OhquR_j88C&oi=fnd&pg=PT8&dq=robust+statistics&ots=rl6VvkJlKU&sig=Su3Fx8s0WfXIb7AapZcADArQzLo&redir_esc=y#v=onepage&q=robust statistics&f=false>.

HUBERT, M., DEBRUYNE, M. (2010). Minimum covariance determinant [online]. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1): 36–43. <https://doi.org/10.1002/wics.61>.

HUBERT, M., DEBRUYNE, M., ROUSSEEUW, P. J. (2018). Minimum covariance determinant and extensions [online]. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(3): 1–11. <https://doi.org/10.1002/wics.1421>.

JOHNSON, R. A. WICHERN, D. A. (2002). *Applied multivariate statistical analysis*. Prentice Hall.

KAPIL, S., CHAWLA, M. (2016). Performance evaluation of K-means clustering algorithm with various distance metrics [online]. *1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, ICPEICES*, pp. 1–4. <https://doi.org/10.1109/ICPEICES.2016.7853264>.

LANTZ, B. (2019). *Machine Learning with R (third)*. Packt Publishing Ltd.

LU, C.-T., CHEN, D., KOU, Y. (2004). Multivariate Spatial Outlier Detection [online]. *International Journal on Artificial Intelligence Tools*, 13(04): 801–811. <https://doi.org/10.1142/S021821300400182X>.

MAULIK, U., BANDYOPADHYAY, S. (2002). Performance evaluation of some clustering algorithms and validity indices [online]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12): 1650–1654. <https://doi.org/10.1109/TPAMI.2002.1114856>.

MILLIGAN, G. W., COOPER, M. C. (1985). An examination of procedures for determining the number of clusters in a data set [online]. *Psychometrika*, 50(2): 159–179. <https://doi.org/10.1007/BF02294245>.

MUSSABAYEV, R., MLADENOVIC, N., JARBOUI, B., MUSSABAYEV, R. (2023). How to Use K-means for Big Data Clustering? [online]. *Pattern Recognition*, 137. <https://doi.org/10.1016/j.patcog.2022.109269>.

NELSON, J. D. (2012). *On K-Means Clustering Using Mahalanobis Distance* [online]. April, pp. 1–113. <https://library.ndsu.edu/ir/bitstream/handle/10365/26766/On K-Means Clustering Using Mahalanobis Distance.pdf?sequence=1>.

PATEL, V. R., MEHTA, R. G. (2012). Data clustering: Integrating different distance measures with modified k-means algorithm [online]. *Advances in Intelligent and Soft Computing*, *131 AISC*, 2: 691–700. <https://doi.org/10.1007/978-81-322-0491-6_63>.

RAEISI, M., SESAY, A. B. (2022). A Distance Metric for Uneven Clusters of Unsupervised K-Means Clustering Algorithm [online]. *IEEE Access*, July, pp. 86286–86297. <https://doi.org/10.1109/ACCESS.2022.3198992>.

SHUKLA, S., LALITHA, S., LALITHA, S. (2021). Spatial Analysis of Water Quality Data Using Multivariate Spatial Outlier Detection Algorithms Spatial data analysis View project Spatial Analysis of Water Quality Data Using Multivariate Spatial Outlier Detection Algorithms [online]. *Ganita*, 70(2): 87–96. <https://www.researchgate.net/publication/369541756>.

SMITI, A. (2020). A critical overview of outlier detection methods [online]. *Computer Science Review*, 38: 100306. <https://doi.org/10.1016/j.cosrev.2020.100306>.

XIE, X. L. BENI, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(8): 841–847.

XU, D., TIAN, Y. (2015). A Comprehensive Survey of Clustering Algorithms [online]. *Annals of Data Science*, 2(2): 165–193. <https://doi.org/10.1007/s40745-015-0040-1>.

XU, R., WUNSCH, D. (2005). Survey of clustering algorithms [online]. *IEEE Transactions on Neural Networks*, 16(3): 645–678. <https://doi.org/10.1109/TNN.2005.845141>.

ZHANG, Z., FENG, Q., HUANG, J., GUO, Y., XU, J., WANG, J. (2021). A local search algorithm for k-means with outliers [online]. *Neurocomputing*, 450: 230–241. <https://doi.org/10.1016/j.neucom.2021.04.028>.

ZHAO, Y., KARYPIS, G. (2002). *Evaluation of Hierarchical Clustering Algorithms for Document Datasets* [online]. Technical Report, pp. 515-524. <https://conservancy.umn.edu/server/api/core/bitstreams/f3a33f1b-eee3-47d5-8cc5-4620054e2ede/content>.