International Journal of Advances in Data and Information Systems
Vol. 7, No. 2, August 2026, pp. 801~815
ISSN: 2721-3056, DOI: 10.25008/ijadis.v7i2.1568

Robust Clustering Analysis of Village Potential Data in South Sumatra Using MMCD-Based Mahalanobis Distance

Dwi Fitrianti[1*], Anwar Fitrianto[2], Anang Kurnia[3]
[1,2,3] School of Data Science, Mathematics, and Informatics, IPB University, Indonesia

Article Info
Article history:
Received Mar 12, 2026
Revised Apr 27, 2026
Accepted July 03, 2026

Keywords:
Robust Mahalanobis Distance
MMCD
Clustering Analysis
Village Potential Data
Outliers

ABSTRACT
Village potential data often contain outliers that can distort clustering results when classical distance measures are applied. This study evaluates the effectiveness of the Robust Mahalanobis Distance based on the Matrix Minimum Covariance Determinant (RMD-MMCD) for clustering village potential data in South Sumatra Province, Indonesia. The proposed approach is compared across K-means, K-medoids, DBSCAN, and DB-Kmeans algorithms, with performance assessed using the Silhouette coefficient, Dunn index, Davies–Bouldin index, Rand Index, and Adjusted Rand Index. The findings demonstrate that incorporating robust covariance estimation through MMCD improves clustering stability and enhances resistance to outliers while preserving the inherent matrix structure of multivariate data. Notable differences in cluster composition are observed primarily in the accessibility and transportation and public service dimensions. These results confirm the advantage of robust distance measures for producing more reliable and structurally consistent clustering solutions in village potential data.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Dwi Fitrianti,
Statistics and Data Sciences, School of Data Science, Mathematics, and Informatics,
IPB University,
IPB University, Dramaga Campus, Bogor, West Java, Indonesia.
Email: dwifitrianti@apps.ipb.ac.id

1. INTRODUCTION

Outliers are one of the challenges that often occur in survey data, including in Village Potential Data (PODES) [1]. Measurement errors, incorrect data entry, and respondents' misinterpretation of questions are all causes of outliers in survey data. Other factors such as social bias and inconsistent respondent responses also contribute to the emergence of outliers [2]. Outliers in the PODES data requires attention, because sustainable national development is highly dependent on the accuracy of regional data. According Badan Pusat Statistik (2024), PODES is data that includes economic, social, and infrastructure facilities which will later be used in program evaluation and the preparation of regional-based policies or strategies. Outliers in PODES data will affect the results of regional grouping and the quality of policies taken. One of the regions that has major challenges related to outliers is South Sumatra Province. According to the publication of the BPS Provinsi Sumatera Selatan (2024), there are 3,283 administrative areas at the village/sub-district level which have various geographical characteristics, ranging from tidal water areas to highlands. The complexity of these village characteristics requires clusterization with a proper cluster analysis so that local government policies can be implemented specifically.

Clustering is a data mining technique that is used to identify clusters without relying on prior knowledge, and can reveal hidden modes and internal structures in a dataset [5]. The main purpose of clustering is to group similar samples and separate different samples into different groups [6]. Various clustering methods have been developed to increase the effectiveness of data clustering.

In the clustering method, the distance generally used is the Euclidean distance, where this distance is not yet resistant to differences in scale and the presence of outliers [5]. Another distance alternative is to use the Mahalanobis distance, which is effective if the data has correlations or features with very different scales, but Mahalanobis itself is still sensitive to outliers because it relies on estimating the covariance of the entire data [7]. Meanwhile, the robust Mahalanobis distance is an extension of the classical Mahalanobis distance constructed using robust estimators of the location and covariance parameters, making it more resistant to the influence of outliers and extreme values [8]. One of the methods frequently used in calculating the robust Mahalanobis distance is the Minimum Covariance Determinant (RMD-MCD), which works by selecting a sub-sample of data that has the smallest covariance matrix determinant [7]. However, RMD-MCD is less than optimal for data with a natural matrix structure because it does not consider the relationships between rows and columns. In addition, the frequently required vectorization process can increase computational complexity and potentially lose important information. As an alternative solution, the robust Mahalanobis distance with Matrix Minimum Covariance Determinant (RMD-MMCD) was developed to calculate the covariance directly from the data in matrix form without the need for vectorization, making it more computationally efficient, preserving the original structure of the data, and more robust against outliers [9]. Based on this, this study modifies the clustering algorithms, namely K-means, K-medoids, DBSCAN, and DB-Kmeans by replacing the distance that generally uses Euclidean distance with RMD-MMCD. This study aims to compare the performance of RMD-MMCD on several clustering methods and analyze the results of regional clustering based on the 2024 District Village Potential (PODES) data in South Sumatra Province.

2. RESEARCH METHOD

This study uses R studio software version 4.5.1 to conduct data analysis with the analysis stages as seen in Figure 1.

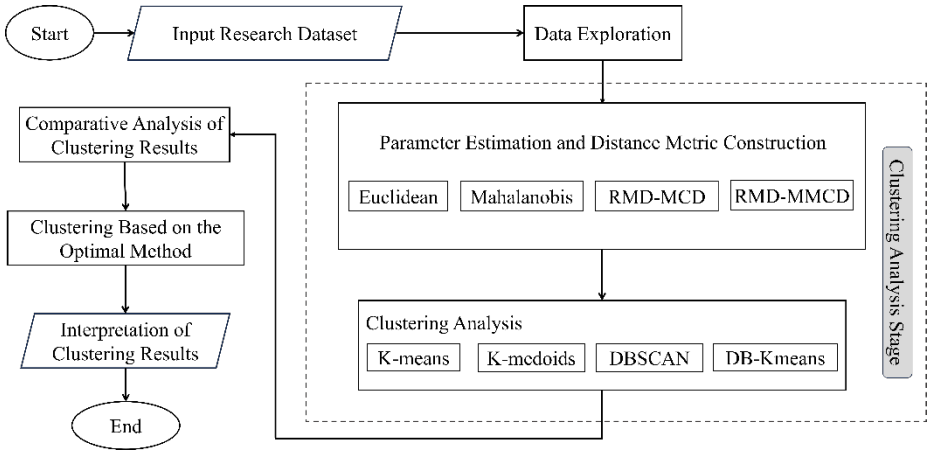


Figure 1. Flowchart of research stages

In this study, the Village Potential (PODES) data of South Sumatera Province in 2024 will be used with a flow diagram of the research stages shown in Figure 1 with the following details:

1. Enter research dataset.
2. Perform data exploration to check for data scale, outliers, and correlations.
3. Estimate the location parameters and covariance classically for Mahalanobis distance (MD) and robustly using MCD and MMCD methods for RMD-MCD and RMD-MMCD. Based on these estimation results, four distance metrics were developed Euclidean distance, MD, RMD-MCD, and RMD-MMCD.
4. Perform cluster analysis by applying K-means, K-medoids, DBSCAN, and DB-Kmeans methods using predefined distance metrics. These methods were selected to represent different clustering approaches, namely partition based methods such as K-means and K-medoids,

density based methods such as DBSCAN, and hybrid approaches such as DB-Kmeans. It should be noted that DBSCAN is able to identify noise points that are not assigned to any cluster, while partition-based methods assign all observations into clusters.
This fundamental difference is acknowledged in this study. Therefore, the comparison is not intended to assume equivalence among methods, but rather to evaluate their performance on the same dataset based on their respective characteristics. In addition, this study aims to evaluate the extent to which the use of robust distance metrics can improve the performance of clustering methods that are not specifically designed to handle outliers.

5. Evaluate the comparison of clustering results carried out to assess the performance of distance metrics on each clustering method using Silhouette, Dunn Index, DBI, Rand Index, and ARI. For DBSCAN, observations identified as noise are excluded from the calculation of evaluation metrics. This is intended to ensure that the comparison remains fair and is not biased by differences in clustering mechanisms across methods.
6. Determining the best clustering method based on the evaluation results, and subsequently performing clustering using the selected method.
7. Interpret the clustering results and draw conclusions.

2.1. Data Types and Sources

This study uses the 2024 Village Potential Data (PODES) of South Sumatra Province. Ten variables with five dimensions on a numerical scale were used, with a total of 3283 villages/sub-districts being observed. The selection of research variables includes five dimensions based on the Village Development Index (IPD) book [10]. The variables used in each dimension in this study can be seen in Table 1.

Table 1. Description of research variables

Dimensions	Variable	Description	Scale
Basic Services	X_1	Total number of educational facilities	Numeric
	X_2	Total number of health workers	Numeric
Infrastructure Conditions	X_3	Total number of economic infrastructure facilities	Numeric
	X_4	Number of Base Transceiver Stations (BTS)	Numeric
Accessibility and Transportation	X_5	Travel time from the village head's office to the sub-district office (minutes)	Numeric
	X_6	Travel time from the village head's office to the regent's office (minutes)	Numeric
Public Services	X_7	Number of sports facilities/fields	Numeric
	X_8	Number of poverty certificates issued by villages in 2023	Numeric
Government Administration	X_9	Number of village government officials	Numeric
	X_{10}	Number of village/sub-district deliberations held in 2023	Numeric

As presented in Table 1, the indicators in each dimension of the Village Development Index (IPD) are structured into multivariate and matrix-variate data. In this study, multivariate data were used to compute Euclidean, Mahalanobis, and RMD-MCD distances, whereas matrix-variate data were structured as a 5×2 matrix for RMD-MMCD estimation. This matrix formulation preserves the original structure of the IPD by retaining the variable pairing within each dimension, thereby maintaining the underlying data characteristics throughout the clustering process..

2.2. Outliers

An outlier is an observation that is located far from the majority of observations in a data set, so that it can affect the overall analysis [11]. In the context of cluster analysis, an outlier refers to data that differs significantly from the pattern or behavior of the majority of other data in a group or cluster. In clustering methods, outliers can affect the distribution of data and interfere with the formation of representative clusters, which can reduce the effectiveness and accuracy of clustering results [12].

2.3. K-means

K-means is a method that attempts to divide data into two or more clusters using the mean value as the cluster center (centroid). This method was introduced in 1967 by MacQueen [13]. The

advantages of K-means lie in its computational efficiency and simplicity, making it widely adopted in large-scale applications. However, this algorithm is sensitive to outliers and initial centroid selection, which can affect the quality of clustering [14]. The K-means clustering procedure is outlined as follows [15] :

1. Determine the number of clusters *(K)*
2. Determine the initial centroids
3. Compute the Euclidean distance between each observation and all centroids using the following equation:

$$d(x,y) = |x - y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

   In this equation, $d(x,y)$ represents the distance between points *x* and *y*, *x* denotes the cluster centroid, *y* refers to the observation of the attribute being measured, *i* indicates each individual observation, *n* is the total number of observations, x_i represents the centroid of the *i*-th cluster, y_i denotes the *i*-th data point.
4. Assign each observation to the cluster with the nearest centroid based on the computed distances.
5. Update the centroids by computing the mean of the observations within each cluster using the following equation:

$$C(i) = \frac{x_i + x_2 + \cdots + x_n}{\sum x} \quad (2)$$

   In this equation, $C(i)$ denotes the updated centroid for the *i*-th cluster, x_n represents the *n*-th observation assigned to that cluster, $\sum x$ refers to the total number of observations within the cluster.
6. Repeat steps 2 through 5 until the cluster centroids achieve convergence.
   The clustering process is considered complete once the centroids achieve convergence or the maximum number of iterations has been reached.

2.4. K-medoids

K-medoids was developed by Kaufman and Rousseeuw in 1987 as an alternative to K-means for handling datasets containing outliers. This method uses actual data points (medoids) as cluster centers, making it more robust to outliers [16]. The algorithm frequently used in K-medoids is Partitioning Around Medoids (PAM). However, this method incurs higher computational complexity compared to K-means, particularly when applied to large-scale datasets. The procedural steps for the K-medoids algorithm are as follows [17]:

1. Determine the number of clusters (K)
2. Randomly select the initial medoids as cluster centers.
3. Compute the Euclidean distance between each observation and the medoids using Equation 2.
4. Identify the nearest medoid for each observation and compute the total cost.
5. Assign each observation to its corresponding medoid to form temporary clusters.
6. Select a new non-medoid object and swap it with the current medoid to evaluate a new set of medoids.
7. Compute the total deviation *(S)* by calculating the difference between the new total cost and the previous total cost.
8. Evaluate the deviation by seeing if $S > 0$ then the process stops with optimal conditions, but if otherwise $S < 0$ then exchange the medoid with other data and repeat the iteration until $S > 0$.

2.4. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a prominent clustering algorithm within density-based approaches [18]. This algorithm was introduced by Ester et al. in 1996 to identify clusters based on local data density and does not require a predetermined number of clusters. [19]. This algorithm is capable of detecting clusters with arbitrary shapes and effectively handling noise. Despite its flexibility, its performance

is highly dependent on precise parameter tuning. The procedural steps of the DBSCAN algorithm are as follows:

1. Define the parameters MinPts (number of neighbors) and Epsilon (distance radius)
2. Identify core points by computing the Euclidean distance to determine if an observation has at least MinPts neighbors within the Epsilon radius.
3. Expand the cluster by identifying density-reachable points from the core points to form high-density regions.
4. Classify border points, which are points located within the epsilon radius of a core point but do not satisfy the MinPts requirement.
5. Identify noise points, which are points that are not reachable from any core point and fail to meet the density criteria.
6. The process terminates once all points have been assigned to a cluster or labeled as noise.

2.5. *DB-Kmeans*

According to research that has been conducted by Gao et al. [20], DB-Kmeans is a method that combines DBSCAN and K-Means to complement each other's shortcomings. This algorithm adopts the cluster center selection dynamically like DBSCAN and reduces the sensitivity of K-means to noise. DB-Kmeans is highly effective in identifying clusters with varying densities and handling high-dimensional data [21]. The main step in the DB-Kmeans method is to perform initial clustering using DBSCAN, which will obtain the number of clusters and their centroids. These outputs will be used as the number of clusters to be formed (K) and the initial centroid. The clustering process will then proceed using the K-means algorithm.

2.6. Mahalanobis Distance

Mahalanobis distance in clustering is utilized to measure the similarity between points within the feature space. This metric is particularly effective for datasets with correlated features or variables of varying scales. However, it remains highly sensitive to outliers, which can distort the mean and covariance estimates, leading to suboptimal clustering results [22]. Empirically, the Mahalanobis distance is computed using the estimated mean and covariance of the dataset as follows:

$$MD(x_i) = \sqrt{\left((x_i - \hat{\mu})^T \hat{\Sigma}^{-1} (x_i - \hat{\mu})\right)} \tag{3}$$

with

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \qquad \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \hat{\mu})\,(x_i - \hat{\mu})^T \tag{4}$$

In this equation, $MD(x_i)$ represents the squared mahalanobis distance from object x_i to the mean value $\hat{\mu}$, x_i denotes the vector of the i -th data observations, $\hat{\mu}$ refers the mean vector and $\hat{\Sigma}$ is the covariance matrix, and n indicates the total number of observations. Mahalanobis distance measures the distance between a point and a distribution by considering the covariance matrix, which effectively accounts for correlations and eliminates redundant information among variables [23].

2.7. Robust Mahalanobis Distance with Minimum Covariance Determinant (RMD-MCD)

The Robust Mahalanobis distance is a modification of the Mahalanobis distance designed to reduce the impact of outliers in distance calculations [24]. MCD is one of the robust methods that is widely used for estimating covariance and mean in the presence of outliers, which provides a robust solution by minimizing the covariance determinant of the most representative subset of data [7]. The MCD estimator is determined by:

$$H^* = \arg\min_{H:|H|=h} \det\left(\widehat{Cov}(X_H)\right) \tag{5}$$

To estimate the covariance matrix using the MCD method:

$$\hat{\mu}_{MCD} = \frac{1}{h}\sum_{i\in H^*}^{h} x_i, \quad \hat{\Sigma}_{MCD} = \frac{1}{h-1}\sum_{i\in H^*}^{h} (x_i - \hat{\mu}_{MCD})\,(x_i - \hat{\mu}_{MCD})^T \tag{6}$$

Thus, the robust mahalanobis distance with MCD is defined as follows:

$$RMD(x_i) = \sqrt{\left((x_i - \mu_{MCD})^T \sum_{MCD}^{-1} (x_i - \mu_{MCD})\right)} \tag{7}$$

In this equation, H^* represents the optimal subset of observations that minimizes the covariance determinant, h denotes the size of the data subset used for estimation, x_i is the i-th observation data vector, $\hat{\mu}_{MCD}$ is the mean MCD dan $\sum_{MCD}^{-1}$ is the MCD covariance, $RMD(x_i)$ is the robust mahalanobis distance with MCD [25].

2.8. Robust Mahalanobis Distance with *Matrix Minimum Covariance Determinant* (RMD-MMCD)

Matrix Minimum Covariance Determinant (MMCD) is an extension of the MCD estimator specifically developed for matrix-valued data. MMCD was introduced by [9] to address the limitations of classical MCD, this approach avoids the need for vectorization. In conventional MCD, matrix data is often vectorized, a process that tends to disregard the inherent row and column structures, leading to less efficient covariance estimation. MMCD provides a more robust and statistically efficient solution than MCD within the context of matrix-variate data.

In the MMCD framework, each observation is represented as a matrix $X_i \in \mathrm{R}^{pxq}$, and the data is modeled using a normal or matrix-elliptical distribution with a Kronecker covariance structure, namely $\Sigma = \Sigma_{col} \otimes \Sigma_{row}$. This structure facilitates the separate modeling of row and column dependencies without requiring vectorization, thereby preserving the inherent structural information of the data. MMCD aims to produce estimators of the central matrix and row and column covariance matrices that are robust to outliers. Robustness is achieved through a trimming mechanism by selecting a subset of observations of size h that is assumed to represent the majority of the data that is free from outliers. The value of h is bounded by $\frac{n}{2} \le h \le n$ and must satisfy $h \ge \left[\frac{p}{q} + \frac{q}{p}\right] + 2$ so that the row and column covariance estimators are positive definite. Based on this subset, MMCD determines the estimator by minimizing an objective function that combines the determinants of the row and column covariances, specifically:

$$\widehat{M}_{H*}, \widehat{\sum}_{row,H*}, \widehat{\sum}_{col,H*} = arg \min_{H\subset\{1,\ldots,n\},|H|=h} \left(p\ \ln det\left(\widehat{\sum}_{col,H}\right) + q\ \ln det\left(\widehat{\sum}_{row,H}\right)\right) \tag{8}$$

By selecting the subset that minimizes the function, a more stable estimator is obtained that is robust to outliers. If $p = 1$ or $q = 1$, then the estimator reverts to MCD on multivariate data. Based on the robust estimator generated through MMCD, a distance measure for matrix-shaped data is defined using the Mahalanobis Matrix distance, namely:

$$MMD\left(X, M, \sum^{row}, \sum^{col}\right) = \sqrt{tr\left(\Omega^{col}(X - M)'\Omega^{row}(X - M)\right)} \tag{9}$$

In this equation, $MMD\left(X, M, \sum^{row}, \sum^{col}\right)$ denotes the robust distance of Mahalanobis Matrix with MMCD, X represents the data matrix, M is the center matrix, $\sum^{row}$ is the row covariance matrix, $\sum^{col}$ is the column covariance matrix, $tr(.)$ is the trace function, Ω^{col} dan Ω^{row} are the inverse row and column covariance matrices. In summary, the MMCD implementation procedure in this study can be formulated as follows:

1. Data structuring, each observation is represented as a matrix of size $p \times q$, and the data are arranged into a three dimensional array of size $p \times q \times n$, and n is the number of observations.
2. Determination of trimming parameter h The trimming parameter $h$is computed using:

$$h = \left[\frac{n + d + 2}{2}\right] \tag{10}$$

where

$$d = \left[\frac{p}{q} + \frac{q}{p}\right] \tag{11}$$

ensuring that $\frac{n}{2} \leq h \leq n$ and that the covariance estimators remain positive definite.

3. Initial subset selection, select an initial subset $H$of size h randomly from the observations.
4. Initial parameter estimation, estimate the center matrix (M) and the row and column covariance matrices (Σ_{row}, Σ_{col}) based on subset H.
5. C-step iteration, iteratively update subset $H$by selecting $h$observations that minimize the objective function until convergence is achieved.
6. Final robust estimation, obtain the robust estimators of the center matrix (M), row covariance (Σ_{row}), and column covariance (Σ_{col}).
7. RMD-MMCD computation, compute the Mahalanobis Matrix Distance between observations using the robust estimators.

2.9. *Davies-Bouldin Index* (DBI)

DBI is a metric used to evaluate the quality of clustering. According to Mughnyanti *et al*. [26], it measures how well a cluster is formed by comparing the intra-cluster distance, or in other words, how close the cluster members are to each other, and the inter-cluster distance, or in other words, how far different clusters are from each other. The lower the DBI value, the better the resulting clustering. The Davies-Bouldin Index is computed using the following formula [27]:

$$DB = \frac{1}{N}\sum_{i=1}^{N} max_{j \neq I}(R_{ij}) \tag{12}$$

In this equation, R_{ij} represents the Davies-Bouldin Index value between clusters C_i and C_j and N indicates the total number of clusters

2.10. *Rand Index* (RI)

Rand Index is a metric utilized to evaluate the quality of clustering results by comparing the generated clustering results with the expected clustering results or ground truth [28]. The Rand Index measures the similarity between two clusters by considering the number of pairs of objects that are in one cluster and in the other cluster are also in the same cluster and vice versa. Menurut Luna-Romera *et al.* [29] The Rand Index is measured by the following equation:

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

In this equation, True Positive (TP) denotes the number of observation pairs assigned to the same cluster in both the generated partitions and the ground truth. True Negative (TN) represents the number of pairs placed in different clusters in both results. Conversely, False Positive (FP) refers to the number of pairs that are in the same cluster in the generated result but in different clusters in the ground truth. False Negative (FN) indicates the number of pairs assigned to different clusters in the generated result but belong to the same cluster in the ground truth. The Rand Index value ranges from 0 to 1, where a value of 1 signifies that the two clusterings are identical, while a value of 0 indicates a total lack of agreement between the two partitions.

2.11. *Adjusted Rand Index* (ARI)

The Adjusted Rand Index evaluates the similarity between two clusters by taking into account the expected index of the Rand Index to correct for bias that may occur due to random similarity [30]. The Adjusted Rand Index is computed according to the following equation:

$$ARI = \frac{RI - Expected\ RI}{Max\ RI - Expected\ RI} \tag{14}$$

In this equation, RI denotes the observed Rand Index, Expected RI represents the expected value of the Rand Index under a random cluster assignment, and Max RI signifies the theoretical maximum value of the Rand Index [31]. The calculations for Expected RI and Max RI are based on the following expressions:

$$Expected\ RI = \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}^2} \tag{15}$$

$$Max\ RI = \frac{1}{2}\left(\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right) \quad (16)$$

In this equation, a_i represents the number of objects in cluster i of the first cluster, and b_j denotes the number of objects in cluster j of the second cluster [32].

2.12. *Silhouette Coefficient*

The silhouette method is a method for determining the optimal number of clusters, the K value with the highest average silhouette score is the best number of clusters. This method assigns each data item a silhouette value, a measure that indicates whether the data is in the correct cluster or closer to another cluster [33]. The silhoutte method has limitations for non-linear clustering, but its widespread use demonstrates its flexibility in a variety of clustering applications [34]. The Silhouette method is used to see the quality and strength of a cluster, how well an object is placed in a cluster with the following equation [35] :

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (17)$$

In this equation, $a(i)$ represents the average distance between object i and all other objects in a cluster, and $b(i)$ denotes the average distance between observation i and all objects in its nearest neighbor cluster. The value of $s(i)$ range between -1and 1, where values close to 1 indicate that the data point fits perfectly with its cluster, values close to 0 indicate that the data point is on the boundary between two clusters, values close to -1 indicate that the data point might fit better if grouped in a different cluster.

2.13. *Dunn Index*

Dunn Index i is an evaluation metric that computes the ratio between the distance between the minimum clusters and the distance within the maximum cluster in a partition [36]. This index serves to assess the density within clusters and the degree of separation between them. Essentially, the Dunn Index aims to identify clusters that are compact and situated as far apart as possible from one another [37]. The equation for the Dunn Index is defined as follows [38]:

$$DI = \frac{min_{1\le i<j\le k}\delta(C_i, C_j)}{max_{1\le l\le k}\Delta(C_l)} \quad (18)$$

In this equation, the numerator represents used to measure the minimum inter-cluster distance (δ). This is the minimum distance between points in two different clusters. We want this value to be large (clusters are far apart). The denominator represents used to measure the maximum intra-cluster distance (Δ). This is the maximum distance between two points within the same cluster (cluster diameter). We want this value to be small (clusters are very dense), and K represents the number of clusters.

3. RESULTS AND DISCUSSION

3.1. Descriptive Data Analysis

The 2024 Village Potential Data for South Sumatra Province encompasses17 regencies and 4 cities. This research only focuses on 17 regencies, so the data used is 3023 villages/sub-districts. employed to characterize the variables and provide an overview of the data distribution. A statistical summary of these variables is presented in Table 2 below.

Table 2. Summary of descriptive statistics of variables

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Mean	4,81	6,21	41,36	0,82	25,81	107,97	3,22	53,07	32,65	5,92
Median	4,00	4,00	30,00	0,00	15,00	90,00	3,00	30,00	28,00	5,00
Max	33,00	125,00	1007,00	13,00	270,00	659,00	10,00	1000,00	165,00	88,00
Min	0,00	0,00	0,00	0,00	1,00	2,00	0,00	0,00	6,00	0,00

As indicated in Table 2, significant scale variations exist among the variables. Variable X_4 as the smallest value range, spanning from 0,00 to 13,00, whereas variables X_6 and X_3 possess substantially larger ranges. Variables with larger values will dominate the calculation [39].

Therefore, a method that is robust to differences in scale is needed. In addition, indications of significant outliers were found in several variables, especially in X_3 where the maximum value reached 1007.00 even though the median value was only in the range of 30.00. To address this, further exploration was conducted through univariate outlier detection using boxplots (Figure 2) and multivariate outlier detection employing the Mahalanobis distance.

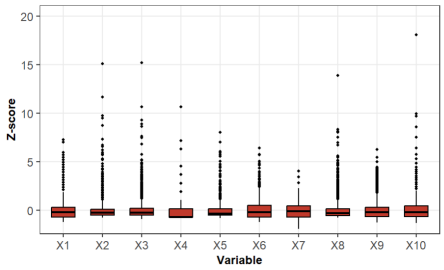


Figure 2. Box plot of all variables used in this study

The results of the detection of univariate outliers produced an average percentage of 5.96% for all variables, and for multivariate outliers using the Mahalanobis distance produced a percentage of 10.22%. In Figure 2, we can also see that all variables contain outliers, which indicates that regional conditions in South Sumatra Province are quite diverse. Following the outlier analysis, the inter-variable correlations were examined, as depicted in the correlation heatmap in Figure 3.

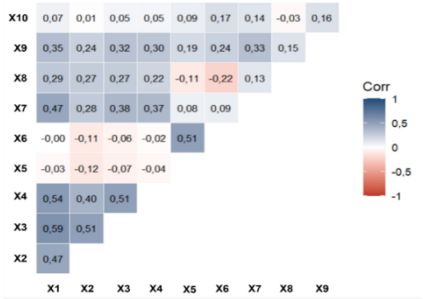
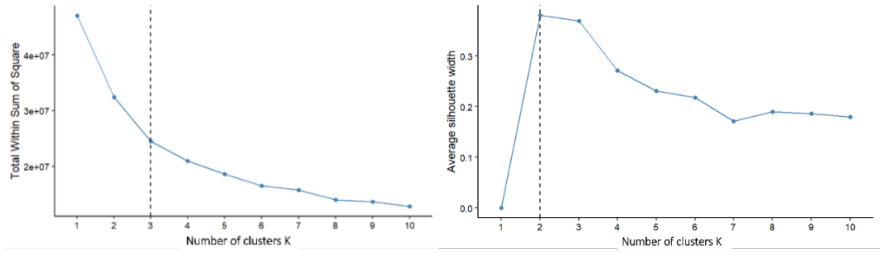


Figure 3. Heatmap of correlation coefficients between variables.

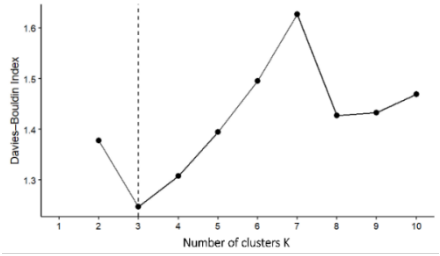
The correlation analysis depicted in Figure 3 indicates that the relationship between variables is dominated by a weak to moderate positive correlation, with the strongest relationship found in the pair X_1 and X_3, which is 0,59. Most of the variables have a low correlation, which indicates that the data is free from serious multicollinearity problems.

3.2. Determining Cluster Parameters

The optimum number of clusters must be determined before clustering is carried out for the K-means and K-medoids methods. In this study, the elbow, silhouette, and DBI methods were used to estimate the optimum number of clusters, as can be seen in Figure 4.



(a) (b)



(c)

Figure 4. Estimation of the optimal number of clusters using (a) the Elbow method, (b) the Silhouette method, and (c) the Davies–Bouldin Index.

The analysis of the Elbow and DBI methods reveals that both metrics converge at K=3. Selecting K=3 is considered more appropriate to accommodate the diversity of village potential data in South Sumatra Province. After the optimum K is obtained for K-means and K-medoids, the next step is to determine MinPts and Epsilon for the DBSCAN parameters. Menurut Salman *et al.* [34] the MinPts value in DBSCAN is generally determined based on the number of variables (d) used in the analysis, with a recommended value range between $d + 1$ to $2d$. In this study, the MinPts value was evaluated using the Silhouette index as an evaluation tool to assess the quality of the formed cluster structure. The epsilon value was determined using the k-Nearest Neighbor (k-NN) distance graph, with the k value adjusted to the selected MinPts. The results of the optimum combination of Minpts and epsilon at each distance can be seen in Table 3 which will later be used in the DBSCAN analysis.

Table 3. Combination of MinPts and Epsilon parameters in the DBSCAN method

Distance	MinPts	Epsilon
Euclidean	11	1,84
Mahalanobis	12	1,68
RMD-MCD	11	2,82
RMD-MMCD	15	2,35

3.3. Performance Analysis of the Clustering Method

Model performance was evaluated using internal validation metrics (Silhouette, Dunn Index, and DBI) and external validation metrics (RI and ARI) by comparing the clustering results with the ground truth obtained from the Village Development Index (IDM). The IDM was used as the ground truth as it is an official government composite indicator aligned with the dimensions of village potential. The initial Village Potential Data (PODES) dataset of South Sumatra Province comprised 3,023 observations. However, since external validation requires the availability of reference labels, only observations with official IDM status were retained for analysis. The reduction in sample size was not due to missing values in the study variables, but rather due to the absence of IDM labels as ground truth. A total of 180 villages/sub-districts without officially released IDM status were excluded from the analysis to ensure the consistent computation of external validation metrics. Consequently, 2,843 observations were retained as the final dataset. Based on the 2024 IDM data, these observations consisted of 183 independent villages, 907 advanced villages, 1,724 developing villages, and 29 underdeveloped villages; thus, the number of clusters was set to K=4. The distribution of IDM labels in this study is imbalanced, particularly for the underdeveloped village category, which has a relatively small number of observations. This condition may affect the Adjusted Rand Index (ARI), as the evaluation results tend to be influenced by classes with larger sample sizes. Therefore, internal validation metrics are also employed to provide a more comprehensive assessment. The evaluation results are presented in Table 4.

Table 4. Results of distance performance evaluation of the clustering method

Distance	Method	Evaluation Metrics				
		Internal			External	
		Sillhoute	Dunn	DBI	RI	ARI
Euclidean	K-means	0,2699	**0,0265**	1,7171	**0,5544**	**0,1077**
	K-medoid	0,2058	**0,0197**	1,8688	**0,5428**	**0,0751**

	DBSCAN	0,4465	0,1167	0,6625	0,5054	**0,0013**
	DB-Kmeans	0,1859	0,0181	1,9110	**0,5260**	**0,0519**
MD	K-means	0,1763	0,0211	2,7706	0,5226	0,0447
	K-medoid	0,0711	0,0160	2,8287	0,5302	0,0459
	DBSCAN	0,3625	**0,1815**	0,8133	0,5113	-0,0017
	DB-Kmeans	0,0851	**0,0229**	3,2824	0,5034	0,0070
RMD-MCD	K-means	0,3057	0,0187	2,0172	0,5138	0,0424
	K-medoid	0,1044	0,0131	2,6361	0,5245	0,0399
	DBSCAN	0,4507	0,1762	0,6664	0,5117	-0,0025
	DB-Kmeans	0,0798	0,0149	3,2714	0,5163	0,0327
RMD-MMCD	K-means	**0,4519**	0,0154	**1,2440**	0,5424	0,1022
	K-medoid	**0,3033**	0,0082	**1,5777**	0,5420	0,0797
	DBSCAN	**0,5290**	0,1699	**0,0973**	**0,5124**	-0,0024
	DB-Kmeans	**0,3611**	0,0116	**1,4658**	0,4953	0,0087

Table 4 demonstrates that the RMD-MMCD distance consistently improves cluster quality across all tested algorithms. The use of RMD-MMCD increases the Silhouette score and reduces the DBI compared to the Euclidean distance. For example, in the DBSCAN algorithm, the RMD-MMCD configuration yields the highest Silhouette score of 0.5290 and the lowest DBI of 0.0973, indicating a compact and well-separated cluster structure. However, external validation metrics such as RI and ARI are relatively small and, in some cases, approach zero or become negative, particularly for the DBSCAN method. This suggests that the resulting cluster structure does not fully align with the reference labels. This discrepancy may be attributed to the fundamental difference between clustering as an unsupervised learning approach, which aims to uncover the natural structure of the data, and the IDM labels, which are constructed based on broader indicators, including non-numerical and policy-related aspects not fully captured in the variables used in this study.

On the other hand, K-means with the RMD-MMCD distance achieves relatively higher ARI values compared to other methods, although it does not always produce the best results in terms of internal validation metrics. This indicates that no single method performs best across all evaluation metrics, and thus the choice of method depends on the objective of the analysis. In line with the objective of clustering as an unsupervised learning approach, greater emphasis is placed on internal validation metrics. This is consistent with the study by by Shutaywi dan Kachouie [40], which states that the Silhouette index does not require a training set to evaluate cluster performance, making it particularly suitable for clustering analysis. However, external validation metrics are still used as complementary measures to assess the agreement between clustering results and reference labels. Overall, the results indicate that the choice of distance metric has a significant impact on clustering performance, particularly for data containing outliers. The RMD-MMCD distance produces more stable and consistent clustering results, making it more suitable for data with outliers and natural matrix-variate structures.

3.4. Regional Clustering Results Based on RMD-MMCD

Table 5 shows a summary of the results of regional grouping based on Village Potential (PODES) data using four clustering methods with RMD-MMCD.

Table 5. Summary of grouping results with MMCD distance

Method	Number of K	*Noise*	Cluster 1	Cluster 2	Cluster 3
K-means	3		2320	281	242
K-medoid	3		777	1724	342
DBSCAN	2	643	2185	15	
DB-Kmeans	2		586	2257	

Table 5 shows that K-means and K-medoids formed three clusters without noise, with K-means producing one dominant cluster, while K-medoids showed a more balanced distribution of members. DBSCAN formed two clusters with the presence of noise, indicating the presence of regions with significantly different PODES characteristics. Meanwhile, DB-Kmeans formed two clusters without noise, indicating that the K-means-based refinement stage was able to regroup the regions so that the cluster structure became more organized. The following discussion is focused on the characteristic profiles of each cluster obtained from the clustering results, as seen in Figure 5.

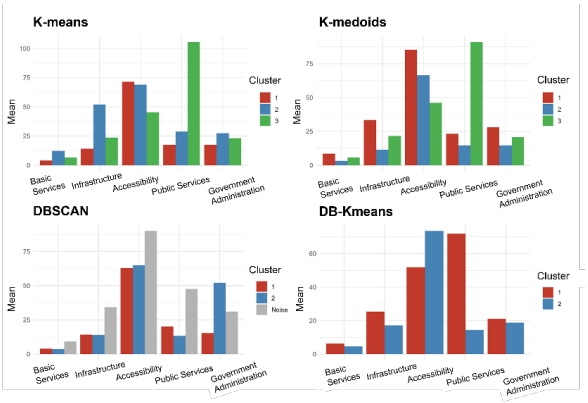


Figure 5. Average profile of village potential dimension scores in each cluster

Figure 5 shows the average profile of village potential dimension scores in each cluster resulting from clustering using K-means, K-medoids, DBSCAN, and DB-Kmeans. In general, cluster differences are more prominent in Accessibility and Transportation (X_5–X_6) and Public Services (X_7–X_8), as indicated by the relatively larger difference in average scores between clusters compared to other dimensions. In Basic Services (X_1–X_2), Infrastructure Conditions (X_3–X_4), and Village Governance (X_9–X_{10}), cluster differences tend to be less pronounced. In DBSCAN, a number of villages were identified as noise with extreme characteristics. This finding indicates that variations in accessibility and public services are the main differentiators between village clusters. Furthermore, a visualization of the map of South Sumatra Province is carried out, which can be seen in Figure 6.

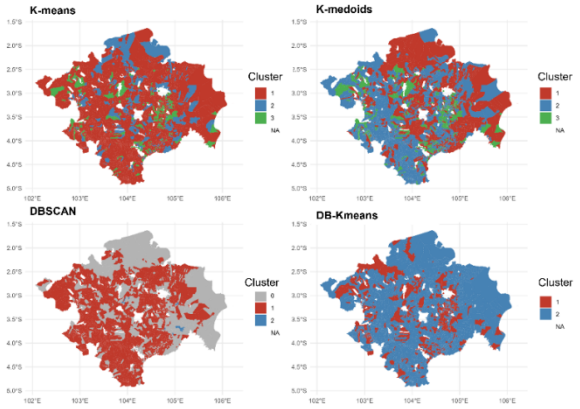


Figure 6. Map of village cluster distribution in South Sumatra Province

Based on Figure 7, the K-means method shows one main cluster dominating most of the area and distributed almost evenly. This pattern indicates that, based on the clustering results, many villages share relatively similar characteristics, while other clusters emerge in smaller pockets. This suggests that K-means tends to capture global similarities without establishing clear spatial relationships.

In contrast to K-means, the K-medoids results display a relatively more structured spatial pattern, with clusters forming more compact and clustered areas. This indicates that differences in village characteristics between clusters are more clearly visible spatially. The DBSCAN method, however, creates one dominant cluster accompanied by scattered noise clusters in certain areas, particularly in areas with characteristics that differ sharply from the surrounding area, in line with DBSCAN's ability to identify low-density or outlier areas. Meanwhile, DB-Kmeans produces two main clusters with a relatively clear regional distribution, making the village clustering pattern simpler and easier to understand.

4. CONCLUSION

The performance of RMD-MMCD on K-means, K-medoids, DBSCAN, and DB-Kmeans clustering methods demonstrated superior performance in handling data containing outliers and with a natural matrix structure compared to other methods. The use of RMD-MMCD was able to reduce the influence of extreme values in distance formation, resulting in a more robust cluster structure and a better representation of the bulk of the data, although the cluster formation mechanisms for each method remained different.

Based on regional clustering using the 2024 District PODES data in South Sumatra Province, regions can be grouped into several clusters with distinct characteristics. Differences between clusters are influenced by the level of regional accessibility and characteristics of public services, while infrastructure and government administration dimensions play a supporting role. These clustering results provide an overview of the diversity of regional characteristics and can be used as supporting information in regional development planning.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support for the publication of this article provided by the Study Program in Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia. The authors also express their sincere gratitude to Anwar Fitrianto and Anang Kurnia for their invaluable guidance, insightful suggestions, and continuous support throughout this research.

CONFLICT OF 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.

DATA AVAILABILITY

The data used in this study were obtained from Statistics Indonesia (BPS) under a data use agreement and are not publicly available. Access to the data is subject to the data access policies and licensing requirements of Statistics Indonesia (BPS).

AUTHOR CONTRIBUTIONS

Dwi Fitrianti: Conceptualization; Data Curation; Formal Analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – Original Draft. **Anwar Fitrianto:** Conceptualization; Methodology; Supervision; Writing – Review & Editing. **Anang Kurnia:** Conceptualization; Methodology; Supervision; Writing – Review & Editing.

REFERENCES

[1] N. Berliana Indah Pratiwi, Indahwati, and A. Fitrianto, "Village Potential Mapping: Comprehensive Cluster Analysis of Continuous and Categorical Variables with Missing Values and Outliers Dataset in Bogor, West Java, Indonesia," *Sci. J. Informatics*, vol. 11, no. 2, pp. 353–366, 2024, doi: 10.15294/sji.v11i2.3903.

[2] E. Zahid, J. Shabbir, and O. A. Alamri, "A Generalized Class of Estimators for Sensitive Variable in the Presence of Measurement Error and Non-Response under Stratified Random Sampling," *J. King Saud Univ. - Sci.*, vol. 34, no. 2, 2022, doi: 10.1016/j.jksus.2021.101741.

[3] Badan Pusat Statistik, "Statistik Potensi Desa Indonesia 2024," Jakarta, 2024. doi: 1105014.

[4] BPS Provinsi Sumatera Selatan, "Statistik Potensi Desa Provinsi Sumatera Selatan 2024," Palembang, 2024. doi: 1105014.16.

[5] T. Li, B. yang Li, X. wei Xin, Y. yuan Ma, and Q. Yang, "A Novel Tree Structure-Based Multi-Prototype Clustering Algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 3, p. 102002, 2024, doi: 10.1016/j.jksuci.2024.102002.

[6] X. Tao *et al.*, "Density Peak Clustering using Global and Local Consistency Adjustable Manifold Distance," *Inf. Sci. (Ny).*, vol. 577, pp. 769–804, 2021, doi: 10.1016/j.ins.2021.08.036.

[7] F. M. Barus and S. Sutarman, "Mendeteksi Outlier pada Data Multivariat dengan Metode Jarak Mahalanobis-Minimum Covariance Determinant (MMCD)," *IJM Indones. J. Multidiscip.*, vol. 1, no. 3, pp. 1164–1172, 2023, [Online]. Available: https://journal.csspublishing/index.php/ijm

[8] R. Lakshmi and T. A. Sajesh, "A Robust Distance-Based Approach for Detecting Multidimensional Outliers," *J. Appl. Stat.*, vol. 52, no. 6, pp. 1278–1298, 2025, doi: https://doi.org/10.1080/02664763.2024.2422403 A.

[9] M. Mayrhofer, U. Radojičić, P. Filzmoser, M. Mayrhofer, and U. Radoji˘, "Robust Covariance Estimation and Explainable Outlier Detection for Matrix-Valued Data Robust Covariance Estimation and Explainable Outlier Detection for Matrix-Valued Data," *Technometrics*, vol. 67, pp. 561–530, 2025, doi: 10.1080/00401706.2025.2475781.

[10] BPS Indoensia, "Indeks Pembangunan Desa 2018," Jakarta, 2018. doi: 1105023.

[11] H. S. Wong and A. Fitrianto, "Split Sample Sequential Fences based on Bootstrap Cut Off Points for Identifying Outliers and Parameter Estimations," *ASM Sci. J.*, vol. 17, no. 3, 2022, doi: 10.32802/asmscj.2022.500.

[12] H. Liu, J. Li, Y. Wu, and Y. Fu, "Clustering with Outlier Removal," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2369–2379, 2021, doi: 10.1109/TKDE.2019.2954317.

[13] W. H. Finch, "A Comparison of Clustering Methods when Group Sizes are Unequal, Outliers are Present, and in the Presence of Noise Variables," *Gen. Linear Model J.*, vol. 45, no. 1, pp. 12–22, 2019, doi: 10.31523/glmj.045001.003.

[14] A. Papadimitriou and V. Tsoukala, "Evaluating and Enhancing the Performance of the K-Means Clustering Algorithm for Annual Coastal Bed Evolution Applications," *Oceanologia*, no. xxxx, 2024, doi: 10.1016/j.oceano.2023.12.005.

[15] J. Medellu and E. S. Nugraha, "K-Medoids Clustering Applications for High-Dimensionality Multiphase," in *Proceeding of The Symposium on Data Science 2021 K-MEANS*, President University Actuarial Science, 2021, pp. 1–13. [Online]. Available: https://e-journal.president.ac.id/index.php/SDS/article/view/1726/965

[16] A. Sobrinho Campolina Martins, L. Ramos de Araujo, and D. Rosana Ribeiro Penido, "K-Medoids Clustering Applications for High-Dimensionality Multiphase Probabilistic Power Flow," *Int. J. Electr. Power Energy Syst.*, vol. 157, no. August 2023, p. 109861, 2023, doi: 10.1016/j.ijepes.2024.109861.

[17] N. F. Fahrudin and R. Rindiyani, "Comparison of K-Medoids and K-Means Algorithms in Segmenting Customers based on RFM Criteria," in *E3S Web of Conferences 484 (FoITIC 2023)*, EDP Sciences, 2024, pp. 1–17. doi: 10.1051/e3sconf/202448402008.

[18] S. S. Li, "An Improved DBSCAN Algorithm Based on the Neighbor Similarity and Fast Nearest Neighbor Query," *IEEE Access*, vol. 8, pp. 47468–47476, 2020, doi: 10.1109/ACCESS.2020.2972034.

[19] A. Latifi-Pakdehi and N. Daneshpour, "DBHC: A DBSCAN-based Hierarchical Clustering Algorithm," *Data Knowl. Eng.*, vol. 135, no. April, p. 101922, 2021, doi: 10.1016/j.datak.2021.101922.

[20] X. Gao, X. Ding, T. Han, and Y. Kang, "Analysis of Influencing Factors on Excellent Teachers' Professional Growth based on DB-Kmeans Method," *EURASIP J. Adv. Signal Process.*, vol. 2022, no. 1, 2022, doi: 10.1186/s13634-022-00948-2.

[21] G. Dong, Y. Jin, S. Wang, W. Li, Z. Tao, and S. Guo, "DB-Kmeans:An Intrusion Detection Algorithm Based on DBSCAN and K-means," *2019 20th Asia-Pacific Netw. Oper. Manag. Symp. Manag. a Cyber-Physical World, APNOMS 2019*, pp. 1–4, 2019, doi: 10.23919/APNOMS.2019.8892910.

[22] P. R. Underhill and T. W. Krause, "Robust Statistics and Cluster Analysis in NDT For Multi-Parameter Signals," *Res. Sq.*, pp. 1–21, 2022, doi: https://doi.org/10.21203/rs.3.rs-1785606/v1.

[23] M. Rafiq, Y. S. Chauhan, and S. Sahay, "Efficient Implementation of Mahalanobis Distance on Ferroelectric FinFET Crossbar for Outlier Detection," *IEEE J. Electron Devices Soc.*, vol. 12, no. May, pp. 516–524, 2024, doi: 10.1109/JEDS.2024.3416441.

[24] Ö. Yorulmaz, S. K. Yıldırım, and B. F. Yıldırım, "Robust Mahalanobis Distance based Topsis to Evaluate the Economic Development of Provinces," *Oper. Res. Eng. Sci. Theory Appl.*, vol. 4, no. 2, pp. 102–123, 2021, doi: 10.31181/oresta20402102y.
[25] M. Hubert and M. Debruyne, "Minimum Covariance Determinant," *WIREs Comput. Stat.*, vol. 2, no. MCD, pp. 36–43, 2010, doi: https://doi.org/10.1002/wics.61.
[26] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of Determining Centroid Clustering X-means Algorithm with Davies-Bouldin Index Evaluation," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, 2020, doi: 10.1088/1757-899X/725/1/012128.
[27] I. T. Umagapi, B. Umaternate, S. Komputer, P. Pasca Sarjana Universitas Handayani, B. Kepegawaian Daerah Kabupaten Pulau Morotai, and B. Riset dan Inovasi, "Uji Kinerja K-Means Clustering Menggunakan Davies-Bouldin Index Pada Pengelompokan Data Prestasi Siswa," *Semin. Nas. SISFOTEK*, pp. 303–308, 2023.
[28] L. Morales and J. Aguilar, "An Automatic Merge Technique to Improve the Clustering Quality Performed by LAMDA," *IEEE Access*, vol. 8, pp. 162917–162944, 2020, doi: 10.1109/ACCESS.2020.3021675.
[29] J. M. Luna-Romera, M. Martínez-Ballesteros, J. García-Gutiérrez, and J. C. Riquelme, "External Clustering Validity Index Based on Chi-Squared Statistical Test," *Inf. Sci. (Ny).*, vol. 487, pp. 1–17, 2019, doi: 10.1016/j.ins.2019.02.046.
[30] S. Dogru and V. Altuntas, "Prediction of Cancer in DNA Sequences using Unsupervised Learning Methods," *J. Innov. Sci. Eng.*, vol. 7, no. 1, pp. 40–47, 2022, doi: 10.38088/jise.1134816.
[31] D. Kallberg, L. Vidman, and P. Ryden, "Comparison of Methods for Feature Selection in Clustering of High-Dimensional RNA-Sequencing Data to Identify Cancer Subtypes," *Front. Genet.*, vol. 12, no. February, 2021, doi: 10.3389/fgene.2021.632620.
[32] Y. Zhu, D. X. Zhang, X. F. Zhang, M. Yi, L. Ou-Yang, and M. Wu, "EC-PGMGR: Ensemble Clustering Based on Probability Graphical Model with Graph Regularization for Single-Cell RNA-seq Data," *Front. Genet.*, vol. 11, no. November, pp. 1–12, 2020, doi: 10.3389/fgene.2020.572242.
[33] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis using Silhouette Score," *Proc. - 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal. DSAA 2020*, pp. 747–748, 2020, doi: 10.1109/DSAA49011.2020.00096.
[34] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *Int. J. online Biomed. Eng.*, vol. 19, no. 4, pp. 174–182, 2023, doi: 10.3991/ijoe.v19i04.37059.
[35] N. Puspitasari, G. Lempas, H. Hamdani, H. Haviuddin, and A. Septiarini, "Perbandingan Algoritma K-Means dan Algoritma K-Medoids Pada Kasus Covid-19 di Indonesia," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, pp. 2015–2027, 2023, doi: 10.47065/bits.v4i4.2994.
[36] H. Malikhatin, A. Rusgiyono, and D. A. Maruddani, "Penerapan K-Modes Clustering dengan Validasi Dunn Index pada Pengelompokan Karakteristik Calon TKI Menggunakan R-GUI," *J. GAUSSIAN*, vol. 10, no. 3, pp. 359–366, 2021, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/
[37] D. Chicco, A. Campagner, A. Spagnolo, D. Ciucci, and G. Jurman, "The Silhouette Coefficient and the Davies-Bouldin Index are More Informative than Dunn index , Calinski-Harabasz Index , Shannon Entropy , and Gap Statistic For Unsupervised Clustering Internal Evaluation of two Convex Clusters," *PeerJ Comput. Sci.*, pp. 1–49, 2025, doi: 10.7717/peerj-cs.3309.
[38] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster Validity Indices for Automatic Clustering : A Comprehensive Review," *Heliyon*, vol. 11, no. 2, p. e41953, 2025, doi: 10.1016/j.heliyon.2025.e41953.
[39] C. Wongoutong, "The Impact of Neglecting Feature Scaling in K-means Clustering," *PLoS One*, vol. 19, no. 12, pp. 1–11, 2024, doi: https://doi.org/10.1371/journal.pone.0310839.
[40] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine," *Entropy*, vol. 23, pp. 1–17, 2021, doi: https://doi.org/10.3390/e23060759.