Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 24, No. 3, July 2025, pp. 579∼590
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v24i3.5057

Comparative Evaluation of Data Clustering Accuracy through Integration of Dimensionality Reduction and Distance Metric

Paska Marto Hasugian[1], Devy Mathelinea[2], Siska Simamora[3], Pandi Barita Nauli Simangunsong[1]
[1]Universitas Katolik Santo Thomas, Medan, Indonesia
[2]Universitas Tun Hussein Onn Malaysia, Johor, Malaysia
[3]Universitas Pembangunan Panca Budi, Medan, Indonesia

Article Info

Article history:

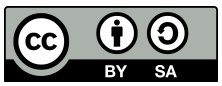
Received May 20, 2025
Revised July 07, 2025
Accepted July 24, 2025

Keywords:

Clustering;
Cluster Evaluation;
Distance Metric;
K-Means;
Principal Component Analysis.

ABSTRACT

The primary issue in clustering analysis of multivariate data is the low accuracy resulting from a mismatch between the Distance Metric used and the characteristics of the data. **This study aims** to comprehensively evaluate the effect of eight Distance Metric in the KMeans algorithm integrated with the Principal Component Analysis (PCA) dimension reduction technique. The analysis process was conducted by transforming the data into two principal components using PCA, then applying K-Means to each Distance Metric. Performance evaluation was conducted based on five internal metrics: Silhouette Score, Davies-Bouldin Index, Sum of Squared Errors, Calinski-Harabasz Index, and Dunn Index. **The results show** that the Bray-Curtis formula provides the best performance, with a Silhouette Score of 0.4291 and SSE of 30.3673. This is followed by Euclidean and Minkowski, which yield the highest Calinski-Harabasz Index value of 2239.85 and Dunn Index of 0.0108, respectively. In contrast, Hamming's formula yielded the lowest performance across all metrics, with a Silhouette Score of 0.0000 and an SSE of 1996.00. The ANOVA test revealed significant differences between the Distance Metric, with a p-value of ¡0.000 for all metrics, which was further supported by the Tukey HSD follow-up test results. **The implications of these findings** confirm the importance of selecting an appropriate Distance Metric in the clustering process to ensure the validity, efficiency, and interpretability of multivariate data analysis results.



Corresponding Author:

Paska Marto Hasugian, +6281264451404,
Faculty of Computer Science, Department of Data Science,
Universitas Katolik Santo Thomas, Medan, Indonesia,
Email: paskamarto86@ust.ac.id

How to Cite:
P. M. Hasugian, D. Mathelinea, S. Simamora, and P. B. N. . Simangunsong, "Comparative Evaluation of Data Clustering Accuracy through Integration of Dimensionality Reduction and Distance Metric", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 24, No. 3, pp. 577-588, July, 2025, doi: 10.30812/matrik.v24i3.5057.

1. INTRODUCTION

The development of digital technology and modern measurement systems has led to the creation of large and high-dimensional data in various sectors, such as health, agriculture, education, and industry [1]. The complexity of the data poses challenges in analysis and exploration, especially in the clustering process [2]. One of the most commonly used algorithms in such a task is K-Means, which works by minimizing the distance between the data and the cluster centers [3], Although popular due to its simplicity, the effectiveness of K-Means is greatly affected by the quality of the data representation and the selection of the Distance Metric used [4]. Scale imbalance between variables, non-normal distribution, the presence of missing values, and the appearance of outliers are some of the factors that can reduce the accuracy and stability of clustering results [5].

Challenges in clustering high-dimensional data, such as differences in scale between variables, uneven data distribution, and the possibility of missing values and outliers, demand an appropriate approach to optimize the cluster structure. One commonly used strategy is to integrate dimensionality reduction techniques and Distance Metric modifications to improve the performance of the K-Means algorithm [6]. Of the various dimensionality reduction techniques, Principal Component Analysis (PCA) is a popular choice as it transforms data to a lower-dimensional space without removing significant variance [7]. The use of Principal Component Analysis (PCA) in the dimensionality reduction process not only improves computational efficiency but also facilitates the revelation of a clearer cluster structure through the representation of data in a simpler feature space. The choice of Distance Metric is also an essential component in the clustering process, as each formula applies a different mathematical approach in measuring the proximity between data, which ultimately affects the quality of clusters, both in terms of internal cohesion and inter-cluster separation [8].

Various studies have been conducted to enhance the effectiveness of data clustering using the K-Means algorithm, both through the selection of a Distance Metric and the application of dimension reduction techniques. [9] Evaluated the effect of various Distance Metrics such as Euclidean, Manhattan, and Chebyshev in the K-Means algorithm and found that the choice of distance greatly affects the clustering results, but did not consider the application of dimension reduction techniques to improve the performance of the algorithm. [10] applied Principal Component Analysis (PCA) to simplify hyperspectral image data before clustering and showed that dimensionality reduction is effective in retaining important information from the data, but this study did not discuss how variations in Distance Metric can affect clustering results after the reduction process. [11] evaluated various distance metrics in the context of medical data and showed the importance of selecting a Distance Metric in producing representative clusters, but without combining it with dimensionality reduction to improve cluster structure. [12] developed a PCA-based environmental quality clustering method with factor weighting, but this study also did not explicitly evaluate the impact of using different Distance Metrics in the clustering results. [13] uses a clustering approach to analyze water quality, considering spatial and distance factors, but does not integrate dimensionality reduction techniques, such as PCA, into its analysis. [14] clustering rice production data with K-Means and Elbow, but only using Euclidean distance without evaluating other distance formulas or dimensionality reduction. [15] in his review discusses various Distance Metrics in the context of data and document clustering, but does not relate them to the use of dimensionality reduction techniques in improving clustering performance.

While these studies have made significant contributions to the development of clustering techniques, most of them either discuss the application of Principal Component Analysis (PCA) and Distance Metrics in isolation or only test certain combinations without thorough exploration. Previous approaches tend to be limited to specific contexts or domains, without presenting a systematic evaluation of the integration of PCA with various Distance Metrics in a uniform experimental framework. The transformation of the feature space due to the application of PCA may affect the way each Distance Metric calculates proximity between data, as each has different sensitivities to the distribution, scale, or orientation of the data. These differences have a direct impact on the structure of the clustering results, especially in terms of internal cluster cohesion and separation between clusters. Thus, a comprehensive approach is needed that not only combines dimensionality reduction techniques and Distance Metric selection but also simultaneously evaluates their impact on clustering quality. This study aims to evaluate the accuracy of K-Means clustering integrated with PCA and various Distance Metrics. The process involves data preparation, dimensionality reduction, and clustering, which are evaluated using the Silhouette Coefficient, Davies-Bouldin Index (DBI), Sum of Squared Errors (SSE), Calinski-Harabasz Index (CHI), and Dunn Index metrics. The results are expected to provide insight into the most effective combination of techniques for clustering high-dimensional data.

2. RESEARCH METHOD

This research employs a mixed-methods approach, combining quantitative and qualitative methods. The research procedure was carried out systematically through several stages to achieve the predetermined objectives. Each stage is designed to ensure the quality of data, the effectiveness of analysis, and the validity of the final results, with the work procedure described in Figure 1.

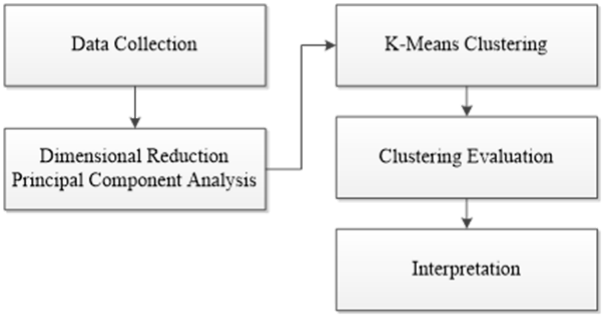


Figure 1. Research work steps

2.1. Data Collection

The dataset used in this study comprises data on the status of villages in Indonesia, totaling 75,248 villages, which can be accessed from the website page https://data.go.id/dataset?q=data. This dataset includes three main variables: the Social Resilience Index (SRI), the Economic Resilience Index (ERI), and the Environmental Resilience Index (ERX). These three indices represent important dimensions in determining the classification of villages as independent, developing, or underdeveloped. This dataset was chosen because it has complex multivariate characteristics, making it suitable for evaluating the performance of clustering methods by integrating dimensionality reduction techniques and various distance formulas, comparatively with the description of the dataset in table 1.

Table 1. Dataset for Clustering Visualization

No	SRI	ERI	ERX	No	SRI	ERI	ERX	No	SRI	ERI	ERX
1	0.8000	0.9000	0.5333	11	0.8514	0.6833	0.6667	75239	0.6686	0.4667	0.5333
2	0.6629	0.6333	0.5333	12	0.7086	0.5500	0.8667	75240	0.4971	0.3667	0.8667
3	0.6743	0.5333	0.6000	13	0.8400	0.9000	0.6667	75241	0.4457	0.2667	0.6667
4	0.7029	0.8000	0.6667	14	0.8857	0.6833	0.8000	75242	0.5657	0.2833	0.9333
5	0.6171	0.6333	0.6000	15	0.8400	0.5833	0.6667	75243	0.6971	0.3667	0.7333
6	0.5771	0.6667	0.6000	16	0.9086	0.7333	0.8000	75244	0.5314	0.2167	1.000
7	0.8114	0.6167	0.6667	17	0.8457	0.6167	0.6000	75245	0.5943	0.5000	0.9333
8	0.8457	0.7833	0.8000	18	0.8400	0.5500	0.6000	75246	0.5714	0.5000	0.7333
9	0.7600	0.6333	0.6667	19	0.7886	0.6333	0.7333	75247	0.7371	0.3000	1.000
10	0.6857	0.6167	0.7333	20	0.7886	0.5833	0.6000	75248	0.5600	0.2500	0.9333

2.2. Dimensionality Reduction with Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is used before the clustering process to reduce the dimensionality of multivariate data, to retain as much information as possible in the form of variance from the original data [16]. This reduction allows the clustering process to run more efficiently and accurately, especially when the data has many variables [17] with the formation of a covariance matrix that represents the linear relationship between variables using the equation 1.

$$c = \frac{1}{n-1} X_{centered}^{T} X_{centered} \tag{1}$$

where C is the covariance matrix, $X_{centered}$ is the data matrix, n is the number of observations, p is the number of variables, which is followed by the calculation of eigen decomposition of the covariance matrix to obtain eigenvalues and eigenvectors with the formula where λ represents the eigenvalue, which indicates the amount of variance explained by the principal component. V is an eigenvector that shows the main direction of the variance. The next step is to determine the relative contribution of each main component to the total variance, as calculated using Equation 2.

$$PC = \frac{\lambda_i}{\Sigma^{\lambda}} \tag{2}$$

Based on the calculation of the variance ratio, components with a cumulative contribution of $\geq 95\%$ were selected, and the data were transformed to a lower-dimensional space using the selected eigenvectors through a linear transformation of $z = XV_k$.

2.3. Clustering with K-Means

K-Means is a clustering algorithm that divides data based on proximity to a centroid, which is chosen randomly and updated iteratively as the average of the data in the cluster until it converges [18]. The similarity between data is measured using various Distance Metric calculation methods, as outlined in equations 3. 4, 5, 6, 7, 8, 9, and 10. Euclidean distance measures the straight distance between points on uniformly scaled continuous data, with the components v_{ik} and v_{jk} of vectors x and y in n dimensions with equation 3 [19]. Manhattan distance calculates the total absolute difference between dimensions and is more robust to outlier values, with the component v_{ik} and v_{jk} of vectors x and y in n dimensions using the equation [20].

$$d(x,y) = \sqrt{\sum_{k=1}^{r}(v_{ik} - v_{jk})^2} \tag{3}$$

$$d(x,y) = \sum_{k=1}^{r}(|v_{ik} - v_{jk}|) \tag{4}$$

Minkowski distance is a generalization of Euclidean and Manhattan distances, by including a parameter of rank p, thus providing flexibility to the shape of the data distribution [21]. Chebyshev distance is used when the largest difference between dimensions is calculated using the equation [22]. Mahalanobis distance considers the covariance between variables and is used in the context of correlated multivariate data [23].

$$d(x,y) = \left(\sum_{k=1}^{r}|v_{ik} - v_{jk}|^2\right)^{\frac{1}{p}} \tag{5}$$

$$d(x,y) = max_i(|v_{ik} - v_{jk}) \tag{6}$$

$$d(x,y) = (x-y)TS - 1(x-y) \tag{7}$$

Canberra distance measures the relative difference between components by using the ratio between the absolute difference and the sum of the values of each component [24]. Bray-Curtis distance calculates dissimilarity based on the ratio of difference to total value and is often used on compositional data [25]. Hamming distance is used for categorical or binary data by calculating the proportion of elements that differ between two vectors, with the following conditions $\delta(x_i, y_i) = 1$ if $x_i \neq y_i$, and 0 if it is equal to the formula [26].

$$d(x,y) = \sum_{i=1}^{n}\frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{8}$$

$$d(x,y) = \frac{\sum_{i=1}^{n}|x_i - y_i|}{\sum_{i=1}^{n}|x_i - y_i|} \tag{9}$$

$$d(x,y) = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i, x_y) \tag{10}$$

2.4. Clustering Evaluation

Clustering evaluation aims to assess how well the cluster results represent the actual data structure [27], with formulas 11, 12, 13, and 14, which are described. Silhouette Score, which measures the extent to which a point fits into its cluster compared to the nearest cluster, where a(i) is the average distance of point i to all points, and b(i) is the average distance. The value of s(i) ranges from -1 to 1, and values close to 1 indicate good cluster separation [28].

$$i = \frac{b(i) - a(i)}{maxa(i), b(i)} \tag{11}$$

Davies-Bouldin Index (DBI) evaluates the quality of clusters based on the compactness and separation between clusters, where s_i and s_j as the average dispersion in the cluster i and j, d_{ij} as the distance between their centroids. The DBI value is the average of the maximum R_{ij} for each i, with the condition that the smaller the value, the better the cluster separation [29].

$$dbi = \frac{1}{k}\sum_{i}^{k} max_{j\neq i}\left(\frac{s_i + s_j}{di_j}\right) \tag{12}$$

Sum of Squared Errors (SSE) is used to measure the internal compactness of the cluster, where x is the data point in the cluster c_i, μ_i is the cluster centroid a small SSE value indicates that the data points are close to the cluster center [30].Calinski-Harabasz Index (CH Index) assesses the ratio of inter-cluster to within-cluster variance by considering the amount of data n and the number of clusters, b_k as the variance between clusters and w_k as the within-cluster variance. High CH values indicate clearly separated clusters [26]. The Dunn Index is used to assess the extent to which clusters are far apart and internally dense. $d(c_i, c_j)$ as the minimum distance between clusters, δ as the maximum diameter in the cluster c_k. A high Dunn Index value indicates a good cluster structure [1].

$$SSE = \sum_{i=1}^{k}\sum_{x\epsilon C_i} ||x - \mu_2||^2 \tag{13}$$

$$CH = \frac{B_k/k - 1}{W_k/(n-k)} \tag{14}$$

$$Di = \frac{min_{i\neq j}}{max_k \triangle (Ck)} \tag{15}$$

3. RESULT AND ANALYSIS

3.1. Dimension Reduction

The dimensionality reduction process is carried out to simplify the data without removing important information needed in further analysis. At this stage, Principal Component Analysis (PCA) was used to reduce the 3 input variables into two principal components. The results of the reduction are used to visualize and analyze data distribution patterns more efficiently, as described in Table 2 below.

Table 2. Reduction with PCA

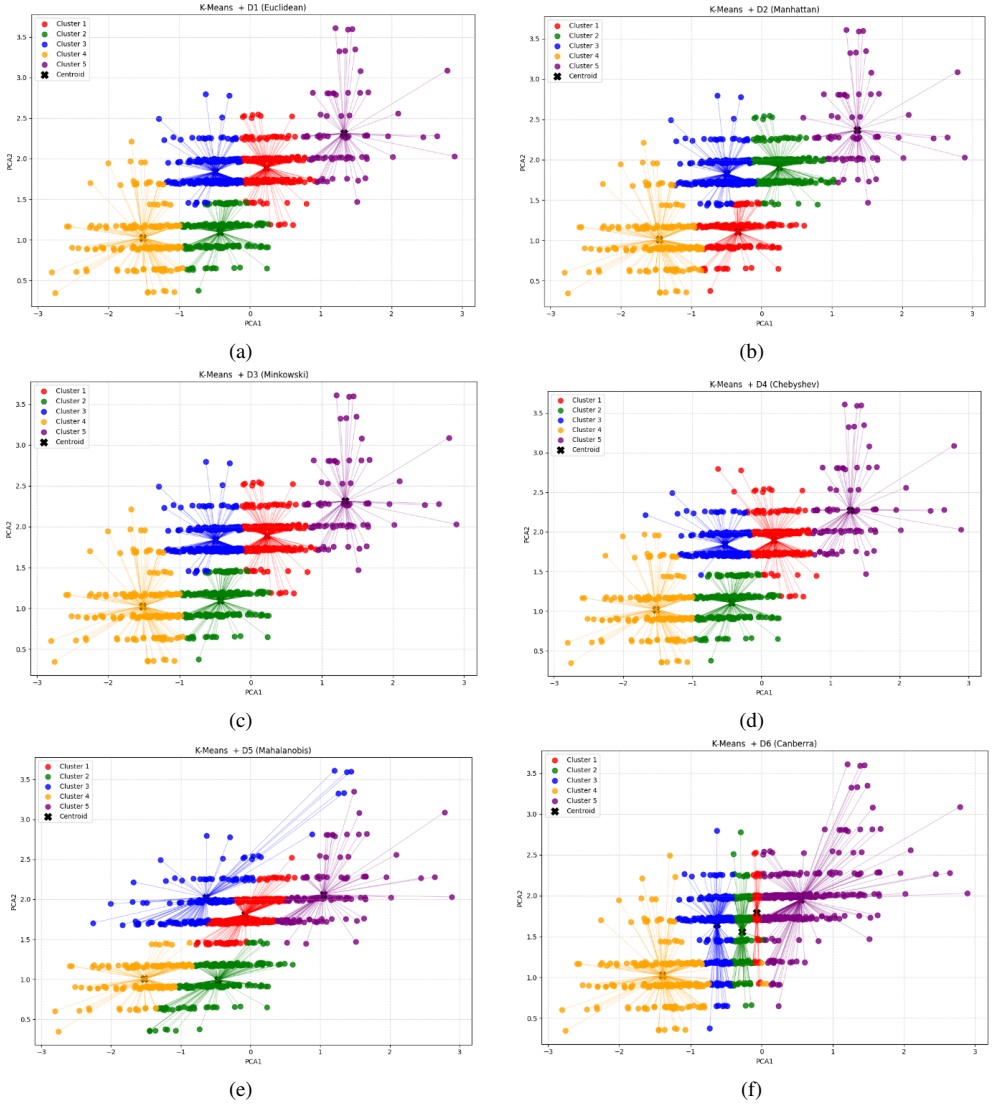
No	PCA1	PCA2	No	PCA1	PCA2	No	PCA1	PCA2
1	-1,08298	2,267578	11	-0,95147	1,724778	75239	2,152862	-0,51251
2	0,542169	2,286493	12	-0,31379	0,938131	75240	2,452366	-1,83298
3	0,650904	2,016999	13	-1,65159	1,724827	75241	3,611497	-1,02044
4	-0,60948	1,7431	14	-1,49114	1,183484	75242	2,214396	-2,10944
5	0,590213	2,023546	15	-0,54416	1,726678	75243	1,808815	-1,32113
6	0,6755	2,028362	16	-1,78163	1,180211	75244	2,437212	-2,3734
7	-0,51659	1,730388	17	-0,50636	1,994304	75245	1,312235	-2,11449
8	-1,63883	1,188156	18	-0,24432	1,995393	75246	1,97773	-1,30616
9	-0,31496	1,73671	19	-0,64237	1,464501	75247	1,106878	-2,40002
10	-0,06465	1,477944	20	-0,10108	2,001614	75248	2,359714	-2,10874

Table 2 presents the results of dimension reduction using Principal Component Analysis (PCA) on the input variables, which are transformed into two principal components: PCA1 and PCA2. Each value in the table represents the coordinates of the data in the new space resulting from the PCA transformation, where each data point is reduced to a pair of values (PCA1, PCA2) that retain most of the information from the original data. These values are used to facilitate the visualization of data in two dimensions, as well as to identify patterns or groups that may be hidden in the original data structure.

3.2. K-Means Clustering with Distance Metric

Clustering is one of the approaches in unsupervised learning that aims to group data based on similar characteristics. In this section, the K-Means method is used as the main clustering algorithm, with variations in the application of various Distance

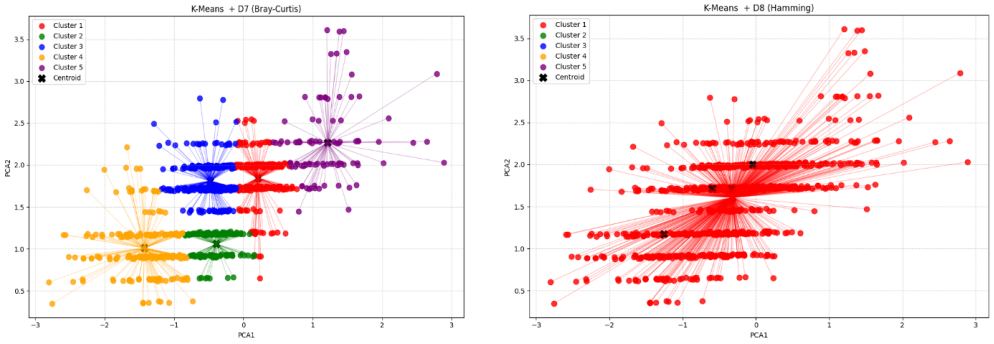
Metric to measure proximity between data. The selection of different Distance Metric is expected to show the effect of distance metrics on the clustering results obtained. The Distance Metric used include Euclidean Distance (D1), Manhattan Distance (D2), Minkowski Distance (D3), Chebyshev Distance (D4), Mahalanobis Distance (D5), Canberra Distance (D6 Bray-Curtis Distance (D7), and Hamming Distance (D8) with clustering results in Figure 2.



(a) (b) (c) (d) (e) (f)

(continued on next page)

Figure 2 (continued)



(g) (h)

Figure 2. Clustering visualization with (a) Euclidean distance (b) Manhattan distance (c) Minkowski distance (d) Chebyshev distance (e) Mahalanobis distance (f) Canberra distance (g) Bray-Curtis distance (h) Hamming distance

After visualizing the cluster distribution based on two-dimensional coordinates (PCA1 and PCA2), further analysis focused on the distribution of each variable within each cluster. The graph in Figure 3 presents the pattern of data distribution for each attribute against the clusters formed based on various Distance Metrics. This visualization aims to identify the internal characteristics of each cluster and clarify the extent to which the different values of the variables form different cluster structures. As such, it provides additional insights into assessing the consistency and quality of cluster formation from the perspective of the underlying attribute values.

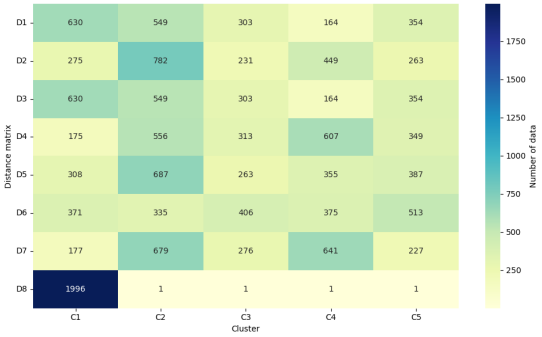


Figure 3. Data distribution against clusters

Figure 3 shows the distribution of the amount of data in each cluster (C1-C5) based on the eight Distance Metric . The Euclidean and Minkowski formulas produce a relatively balanced distribution, with C1 containing 630 data, C2 with 549, C3 with 303, C4 with 164, and C5 with 354. Manhattan tends to center on C2 with 782 data, while Chebyshev dominates C4 with 607 data. The Mahalanobis formula spreads fairly evenly, with peaks at C2 with 687 data and C5 with 387. Canberra and Bray-Curtis show a rather uneven distribution, with Canberra highest at C5 with 513 and Bray-Curtis at C2 and C4 with 679 and 641 respectively. Meanwhile, Hamming failed to form a proper clustering as 1996 data were concentrated in C1 and only 1 data in the other clusters.

3.3. Internal Evaluation of Clustering

Clustering performance evaluation is conducted to measure the extent to which the quality of cluster formation is statistically and interpretatively acceptable. In this study, five evaluation metrics were used, namely Silhouette Score, Davies-Bouldin Index, Sum of Squared Errors (SSE), Calinski-Harabasz Index, and Dunn Index. These five metrics are used to assess the internal coherence of the clusters, the separation between clusters, the degree of spread of the data with respect to the cluster centers, and the density and minimum distance between clusters. With the combination of these metrics, the performance of each Distance Metric in forming clusters can be comprehensively evaluated, both in terms of geometric structure and clustering stability, as depicted in Figure 4.

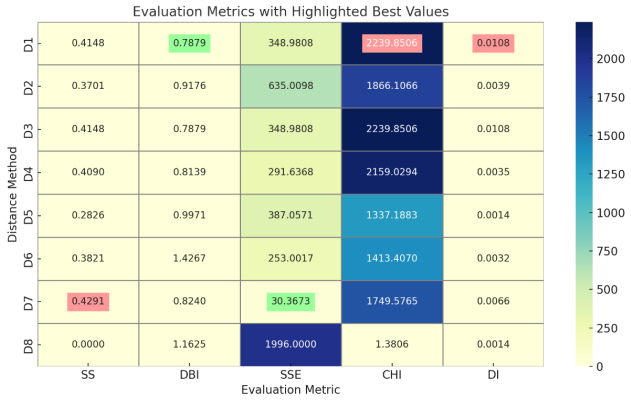


Figure 4. Evaluation metric comparisons

Figure 4 presents an internal evaluation of the clustering, which reveals that the Bray-Curtis Distance Metric yields the best performance, with a Silhouette Score of 0.4291 and a Sum of Squared Errors value of 30.3673, indicating a very compact and well-separated cluster. The Euclidean and Minkowski formulas also performed consistently, with a Silhouette Score of 0.4148, a Davies-Bouldin Index of 0.7879, the highest Calinski-Harabasz Index of 2,239.8506, and the highest Dunn Index of 0.0108, reflecting a clear and efficient cluster structure. In contrast, the Hamming formula showed the worst performance, with a Silhouette Score of 0.0000, a Davies-Bouldin Index of 1.1625, a Sum of Squared Errors of 1996, a Calinski-Harabasz Index of 1.3806, and a Dunn Index of 0.0014. In the context of theory, Hamming distance is a Distance Metric designed for binary categorical data, such as strings or vectors of 0 and 1. It measures the number of different positions between two vectors, without considering the absolute value or scale of the attribute. Therefore, when applied to continuous numerical data, Hamming distance cannot represent geometric relationships between data and tends to produce invalid cluster structures. The use of Hamming distance on this type of data is conceptually inappropriate, resulting in low evaluation metric results that do not accurately reflect the true quality of clustering.

3.4. Statistical Evaluation of Clustering

A statistical evaluation was conducted to compare the performance of the eight distance methods on the clustering results. Five evaluation metrics were used: Silhouette Coefficient, Davies-Bouldin Index (DBI), Sum of Squared Errors (SSE), Calinski-Harabasz Index (CHI), and Dunn Index. Tests were conducted using Analysis of Variance (ANOVA) to assess whether there were statistically significant differences between the distance method groups for each evaluation metric. ANOVA was conducted under the assumptions that the data were normally distributed, had homogeneous variance (homoskedasticity), and the observations were independent. A difference was considered significant if the p value was ¡0.05. This test was followed by Tukey's Honestly Significant Difference (Tukey HSD) as a post-hoc test to identify significantly different method pairs, supported by the 95% confidence intervals outlined in table 3.

Table 3. ANOVA Test Results

Metrik	F Value	p-value
Silhouette	$4.46 \times 10^{3}0$	0.000
DBI	$4.66 \times 10^{3}0$	0.000
SSE	$2.78 \times 10^{3}1$	0.000
CH	$2.28 \times 10^{3}1$	0.000
Dunn	$3.37 \times 10^{3}0$	0.000

The ANOVA test results indicate that the selection of the distance method has a significant impact on all five clustering evaluation metrics: Silhouette Coefficient, Davies-Bouldin Index, Sum of Squared Errors, Calinski-Harabasz Index, and Dunn Index. This is evidenced by the very large F-values (up to the order of 10^{31}) and p-values of 0.000 on all metrics, indicating that the performance differences between methods are not random, but statistically significant at the 95% confidence level. Thus, further analysis using the Tukey HSD post-hoc test was required to identify pairs of distance methods that were significantly different, as described in Table 4.

Table 4. Tukey HSD Test Results Cluster Evaluation Metrics (p = 0.000)

No	Evaluation Metrics	Distance Metrics Pairs	Mean Difference	Description
		BrayCurtis vs Hamming	0.4291	Highest difference
		Euclidean vs Hamming	-0.4148	Negatively significant
1	Silhouette Coefficient	Mahalanobis vs BrayCurtis	-0.1464	Negatively significant
		Manhattan vs Minkowski	0.0448	Significant small positive
		Chebyshev vs Mahalanobis	-0.1264	Significant
		Canberra vs BrayCurtis	-0.6026	Negative highest difference
		Canberra vs Euclidean	0.6388	Highest difference
2	Davies-Bouldin Index	Manhattan vs BrayCurtis	-0.0935	Significant
		Chebyshev vs Hamming	-0.3486	Significant
		Mahalanobis vs BrayCurtis	-0.1730	Significant
		BrayCurtis vs Hamming	-1965.6	The biggest difference
		Chebyshev vs Hamming	-1704.3	Significant
3	SSE	Canberra vs Hamming	-1742.9	Significant
		Mahalanobis vs Hamming	-1608.9	Significant
		Euclidean vs Manhattan	-286.03	Significant
		Euclidean vs Hamming	2238.4	Highest CH value
		Canberra vs Hamming	1412.0	Significant
4	Calinski-Harabasz	Mahalanobis vs Hamming	1335.8	Significant
		Manhattan vs Hamming	1864.7	Significant
		Minkowski vs Hamming	2238.4	Same height as Euclidean
		Minkowski vs Mahalanobis	0.0094	Highest difference
		BrayCurtis vs Hamming	0.0052	Significant positive
5	Dunn Index	Manhattan vs Minkowski	-0.0068	Significant positive
		Canberra vs BrayCurtis	-0.0034	Significant positive
		Chebyshev vs Mahalanobis	-0.0021	Significant positive

According to the Silhouette Coefficient metric, the Bray-Curtis method performed significantly better than Hamming, with the highest average difference of 0.4291. In contrast, Hamming is consistently the lowest-performing method on most metrics, reflected by significant negative differences with other methods, such as Euclidean and Mahalanobis, for Davies-Bouldin Index (DBI), where the lower the value the better, BrayCurtis again outperformed other methods such as Canberra and Mahalanobis, with the largest difference of -0.6026 against Canberra, indicating that BrayCurtis produces more compact and separated clusters. In the Sum of Squared Errors (SSE) metric, the largest difference was recorded between Bray-Curtis and Hamming (-1965.63), indicating that Hamming produces a very high error rate. According to the Calinski-Harabasz Index (CH), the Euclidean and Minkowski methods are significantly superior to the Hamming method, with a difference of up to 2238.47, reflecting more clearly defined and separated clusters. For the Dunn Index, the largest significant difference appeared in the pairing of Minkowski and Mahalanobis (0.0094), confirming the superiority of Minkowski in forming compact and widely separated clusters. These results demonstrate that the selection of an appropriate distance method has a significant impact on the quality of the clustering results. Methods such as Bray-Curtis, Euclidean, and Minkowski more consistently show superior performance, while Hamming proves less suitable for continuous numerical data, as in this case. This finding is reinforced by the presentation of 95% confidence intervals for mean differences between methods, which provide strong statistical evidence for the stability and superiority of a particular method in producing optimal clustering, summarized in Table 5 Confidence Interval (95%)

Table 5. Confidence Interval (95%)

Metrics	Methods	Mean	CI Lower	CI Upper
Silhouette	Bray-Curtis	0.429	0.429	0.429
	Hamming	0.000	0.000	0.000
DBI	Canberra	1.426	1.426	1.426
	Euclidean	0.787	0.787	0.787
SSE	Hamming	1996.00	1996.00	1996.00
	Bray-Curtis	30.37	30.37	30.37
CH	Euclidean	2239.85	2239.85	2239.85
	Hamming	1.38	1.38	1.38
Dunn	Minkowski	0.0108	0.0108	0.0108
	Mahalanobis	0.0014	0.0014	0.0014

The 95% confidence intervals (CIs) show the stability of each distance method's performance, with mean values identical to the lower and upper limits of the CIs. Bray-Curtis performed best in Silhouette (0.429) and SSE (30.37), indicating compact clusters and minimum error. In contrast, Hamming consistently performed poorly, with a Silhouette value of 0.000 and the highest SSE (1996.00). This statistically reflects its inability to measure proximity on continuous numerical data, as Hamming was designed for binary categorical data, not ratio-scale variables. Euclidean performed highly on the CH Index (2239.85) and had a low DBI (0.787), reflecting well-defined clustering. On the Dunn Index, Minkowski excelled with the highest value (0.0108), indicating compact and separated clusters. Overall, the Bray-Curtis, Euclidean, and Minkowski methods yielded the most effective and consistent clustering results, whereas the Hamming method was not suitable for this data context.

This research focuses on the comparative evaluation of data clustering accuracy by integrating the Principal Component Analysis dimension reduction technique with eight variations of distance formulas in the K-Means algorithm. This approach is designed to address methodological limitations in previous studies that generally evaluate the effectiveness of distance formulas or the application of dimensionality reduction separately, without considering their synergy in forming an optimal cluster structure. Performance evaluation is conducted through five internal metrics: Silhouette Score, Davies-Bouldin Index, Sum of Squared Errors, Calinski-Harabasz Index, and Dunn Index, which cover aspects of density, separation, and stability of the cluster structure. Furthermore, to ensure that the performance differences between the method combinations were statistically significant, analysis of variance and Tukey HSD follow-up tests were applied. This combination of evaluative approaches yields an analysis that is not only descriptive but also inferential, thereby providing a deeper and more measurable understanding of the effectiveness of the method integration used. Thus, this research makes a meaningful contribution to the study of data clustering, especially in terms of a more integrated evaluation methodology supported by empirical evidence.

4. CONCLUSION

The internal evaluation of the clusterization shows that the quality of the results is strongly influenced by the Distance Metric used. The Bray-Curtis formula performed best with a Silhouette Score of 0.4291 and the lowest SSE of 30.3673, reflecting compact and well-separated cluster formation. The Euclidean and Minkowski formulas also yielded strong results, particularly in terms of the Calinski-Harabasz Index (2239.85) and Dunn Index (0.0108) metrics, indicating an efficient and well-defined cluster structure. In contrast, Hamming's formula yielded the worst results on all metrics, including a Silhouette Score of 0.0000 and an SSE of 1996.00, indicating its unsuitability for continuous numerical data. Statistical evaluation through ANOVA resulted in very large F-values and p-values of 0.000 on all metrics, indicating that the difference in performance between formulas was statistically significant. The Tukey HSD follow-up test revealed that the performance differences between certain formula pairs, such as Bray-Curtis and Hamming, were statistically significant. The implication is that selecting an appropriate Distance Metric is crucial in the clustering process to ensure valid and interpretable results, especially in large-scale and multivariate data analysis, such as village resilience. Future research may focus on the development of new distance formulations or adaptive metrics that are better suited to the structure of high-dimensional and heterogeneous datasets, as well as their application in other domains requiring robust cluster interpretation.

5. ACKNOWLEDGEMENTS

The authors would like to thank the Department of Computer Science for supporting the computing facilities used in this research. Thanks are also due to the colleagues who provided constructive feedback, language editing assistance, and manuscript correction.

6. DECLARATIONS

AUTHOR CONTIBUTION

Paska Marto Hasugian, Conceptualization, Methodology, Writing - Initial Draft. Devy Mathelinea, Data Curation, Formal Analysis, Visualization. Siska Simamora, Programming, Validation, Writing - Review & Editing. Pandi Barita Nauli Simangunsong, Supervision, Project Administration, Final Approval of Manuscript.

FUNDING STATEMENT

This research has received no specific funding from any public, commercial, or non-profit organization.

COMPETING INTEREST

The authors declare that they have no financial or personal conflicts of interest that could influence the results of this study.

REFERENCES

[1] C.-E. Ben Ncir, A. Hamza, and W. Bouaguel, "Parallel and scalable Dunn Index for the validation of big data clusters," *Parallel Computing*, vol. 102, p. 102751, 2021, https://doi.org/10.1016/j.parco.2021.102751.

[2] S. Suboh, I. A. Aziz, S. M. Shaharudin, S. A. Ismail, and H. Mahdin, "A Systematic Review of Anomaly Detection within High Dimensional and Multivariate Data," vol. 7, no. March, 2023, https://doi.org/10.30630/joiv.7.1.1297.

[3] J. Yin, S. Sun, L. Wei, and P. Wang, "Discriminatively Fuzzy Multi-View K-means Clustering with Local Structure Preserving," vol. 38, no. 5, pp. 16 478–16 485, 2024, https://doi.org/10.1609/aaai.v38i15.29585.

[4] M. Zubair, M. D. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker, "An improved K-means clustering algorithm towards an efficient data-driven modeling," *Annals of Data Science*, vol. 11, no. 5, pp. 1525–1544, 2024, https://doi.org/10.1007/s40745-022-00428-2.

[5] J. Zhao, G. Wang, J.-S. Pan, T. Fan, and I. Lee, "Density peaks clustering algorithm based on fuzzy and weighted shared neighbor for uneven density datasets," *Pattern Recognition*, vol. 139, p. 109406, July, 2023, https://doi.org/10.1016/j.patcog.2023.109406.

[6] O. Dorabiala, A. Y. Aravkin, and J. N. Kutz, "Ensemble principal component analysis," *IEEE Access*, vol. 12, pp. 6663–6671, January, 2024, https://doi.org/10.1109/ACCESS.2024.3350984.

[7] F. Zou and G. G. Yen, "Dynamic multiobjective optimization with varying number of objectives assisted by dynamic principal component analysis," *Information Sciences*, vol. 665, p. 120398, April, 2024, https://doi.org/10.1016/j.ins.2024.120398.

[8] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, vol. 8, no. March, pp. 54 776–54 788, 2020, https://doi.org/10.1109/ACCESS.2020.2980942.

[9] G. K. Patel, V. K. Dabhi, and H. B. Prajapati, "Clustering Using a Combination of Particle Swarm Optimization and K-means," vol. 26, no. 3, pp. 457–469, May, 2017, https://doi.org/10.1515/jisys-2015-0099.

[10] K. Yu, S. Fang, and Y. Zhao, "Heavy metal Hg stress detection in tobacco plant using hyperspectral sensing and data-driven machine learning methods," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 245, p. 118917, 2021, https://doi.org/10.1016/j.saa.2020.118917.

[11] C. E. Coombes, X. Liu, Z. B. Abrams, K. R. Coombes, and G. Brock, "Simulation-derived best practices for clustering clinical data," *Journal of Biomedical Informatics*, vol. 118, p. 103788, June, 2021, https://doi.org/10.1016/j.jbi.2021.103788.

[12] M. Tripathi and S. K. Singal, "Allocation of weights using factor analysis for development of a novel water quality index," *Ecotoxicology and Environmental Safety*, vol. 183, p. 109510, November, 2019, https://doi.org/10.1016/j.ecoenv.2019.109510.

[13] A. C. P. Fernandes, L. F. S. Fernandes, R. M. V. Cortes, and F. A. L. Pacheco, "The role of landscape configuration, season, and distance from contaminant sources on the degradation of stream water quality in urban catchments," *Water (Switzerland)*, vol. 11, no. 10, 2019, https://doi.org/10.3390/w11102025.

[14] P. M. Hasugian, B. Sinaga, J. Manurung, and S. A. Al Hashim, "Best Cluster Optimization with Combination of K-Means Algorithm And Elbow Method Towards Rice Production Status Determination," *International Journal of Artificial Intelligence Research*, vol. 5, no. 1, pp. 102–110, 2021, https://doi.org/10.29099/ijair.v6i1.232.

[15] S. Sumathi and H. G. Gunaseelan, "A Review of Data and Document Clustering pertaining to various Distance Measures," *Salud, Ciencia y Tecnología*, 2022, https://doi.org/10.56294/saludcyt2022194.

[16] N. Faris, A. Sahi, M. Diykh, S. Abdulla, and S. Siuly, "Enhanced Polycystic Ovary Syndrome Diagnosis Model Leveraging a K-means Based Genetic Algorithm and Ensemble Approach," *Intelligence-Based Medicine*, vol. 11, p. 100253, 2025, https://doi.org/10.1016/j.ibmed.2025.100253.

[17] R. Perera, M. C. Huerta, C. Barris, and M. Baena, "Clustering classifier of FRP strengthened concrete beams using superpixels and principal component analysis," *Construction and Building Materials*, vol. 453, no. June, p. 139019, 2024, https://doi.org/10.1016/j.conbuildmat.2024.139019.

[18] A. K. Abdalameer, M. Alswaitti, A. A. Alsudani, and N. A. M. Isa, "A new validity clustering index-based on finding new centroid positions using the mean of clustered data to determine the optimum number of clusters," *Expert Systems with Applications*, vol. 191, p. 116329, April, 2022, https://doi.org/10.1016/j.eswa.2021.116329.

[19] Q. Zhang, X. Zhang, J. Yang, M. Sun, and T. Cao, "Introducing Euclidean distance optimization into Softmax loss under neural collapse," *Pattern Recognition*, vol. 162, no. November 2024, p. 111400, 2025, https://doi.org/10.1016/j.patcog.2025.111400.

[20] Y. Yuan, J. Wang, W. Li, K. Wang, H. Rao, and J. Xu, "Fast supervoxel segmentation of connectivity median simulation based on Manhattan distance," *International Journal of Applied Earth Observation and Geoinformation*, vol. 133, p. 104108, September, 2024, https://doi.org/10.1016/j.jag.2024.104108.

[21] S. Liaquat, M. F. Zia, O. Saleem, Z. Asif, and M. Benbouzid, "Performance analysis of distance metrics on the exploitation properties and convergence behaviour of the conventional firefly algorithm[Formula presented]," *Applied Soft Computing*, vol. 126, p. 109255, September, 2022, https://doi.org/10.1016/j.asoc.2022.109255.

[22] N. Krivulin, "Algebraic solution of minimax single-facility constrained location problems with Chebyshev and rectilinear distances," *Journal of Logical and Algebraic Methods in Programming*, vol. 115, p. 100578, October, 2020, https://doi.org/10.1016/j.jlamp.2020.100578.

[23] A. Ghosh, A. K. Ghosh, R. SahaRay, and S. Sarkar, "Classification Using Global and Local Mahalanobis Distances," vol. 207, no. February 2024, 2024, https://doi.org/10.1016/j.jmva.2025.105417.

[24] P. M. Hasugian, H. Mawengkang, P. Sihombing, and S. Efendi, "Development of distance formulation for high-dimensional data visualization in multidimensional scaling," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 2, pp. 1178–1189, 2025, https://doi.org/10.11591/eei.v14i2.8738.

[25] W. Zhao, L. Yang, C. Dang, R. Rocchetta, M. Valdebenito, and D. Moens, "Enriching stochastic model updating metrics: An efficient Bayesian approach using Bray-Curtis distance and an adaptive binning algorithm," *Mechanical Systems and Signal Processing*, vol. 171, no. September 2021, p. 108889, 2022, https://doi.org/10.1016/j.ymssp.2022.108889.

[26] P. Agarwalla and S. Mukhopadhyay, "Gene expression selection for cancer classification using intelligent collaborative filtering and hamming distance guided multi-objective swarm optimization," *Applied Soft Computing*, vol. 170, no. November 2024, p. 112654, 2025, https://doi.org/10.1016/j.asoc.2024.112654.

[27] J. Wu, J. Chen, H. Xiong, and M. Xie, "External validation measures for K-means clustering: A data distribution perspective," *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 6050–6061, 2009, https://doi.org/10.1016/j.eswa.2008.06.093.

[28] A. Arunkumar, A. Pinceti, L. Sankar, and C. Bryan, "PMU Tracker: A Visualization Platform for Epicentric Event Propagation Analysis in the Power Grid," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1081–1090, jan 2023, https://doi.org/10.1109/TVCG.2022.3209380.

[29] I. K. Khan, H. Daud, N. Zainuddin, and R. Sokkalingam, "Standardizing reference data in gap statistic for selection optimal number of cluster in K-means algorithm," *Alexandria Engineering Journal*, vol. 118, no. January, pp. 246–260, 2025, https://doi.org/10.1016/j.aej.2025.01.034.

[30] M. Raeisi and A. B. Sesay, "A Distance Metric for Uneven Clusters of Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 10, no. August, pp. 86 286–86 297, 2022, https://doi.org/10.1109/ACCESS.2022.3198992.