Daengku: Journal of Humanities and Social Sciences Innovation *ISSN: 2775-6165 (online)*
Vol. 5 No. 4 (2025)
https://doi.org/10.35877/454RI.daengku4439

Comparative Performance Analysis of Two Clustering Methods for Grouping Indonesian Provinces Based on Forest Area Size

Sitti Masyitah Meliyana*, Anugra S.A. Dunggio, Subhan Muhammad, & Abdul Rahman

Department of Statistics, Faculty of Mathematics and Natural Science, Universitas Negeri Makassar, Makassar and 90224, Indonesia

Abstract

This study aims to compare the performance of two clustering algorithms, K-Means Clustering and K-Medoids Clustering in grouping Indonesian provinces based on forest area by type. The optimal number of clusters was determined using the minimum Davies–Bouldin Index (DBI), while cluster performance was evaluated using the Silhouette Coefficient. Clustering, as one of the key techniques in data mining, automatically classifies data into several groups with similar characteristics. The results reveal differences in the number of clusters produced by the two algorithms. The K-Means method generated four clusters, indicated by its lowest DBI value of 0.515, whereas the K-Medoids method produced three clusters, with a minimum DBI value of 0.559. The clustering performance of K-Means resulted in a Silhouette Coefficient of 0.610, while K-Medoids achieved a higher value of 0.644. Based on these results, the K-Medoids Clustering method with three clusters, demonstrates superior performance in analyzing the grouping of Indonesian provinces by forest area type.

Keywords: Forest, Davies–Bouldin Index, K-Means Clustering, K-Medoids Clustering, Silhouette Coefficient

1. Introduction

Forests are vital ecosystems that support terrestrial biodiversity and serve as an essential resource for human life (Pratiwi et al., 2024). Forests absorb carbon dioxide and produce oxygen, making Indonesia one of the world's major "lungs" that contributes significantly to global ecological balance (Shafitri et al., 2018). For local communities, forests play an important role not only as sources of daily necessities but also as areas for economic activity. Sustainable management of non-timber forest products and environmental services is therefore crucial for improving community welfare. However, high dependence on forest resources and competing political and economic interests particularly related to logging and land conversion continue to undermine ecological functions and affect the social conditions of forest dependent communities. In addition, forest degradation continues to increase over time due to rising demands for land that are not matched by sustainable land expansion, resulting in misuse and inappropriate land allocation (Febryanti et al., 2020).

According to data compiled by the University of Maryland and the World Resources Institute, global forest loss reached 41,000 km² in 2022, exceeding the level recorded in 2021 despite global commitments to reduce deforestation (Noer & Dimyati, 2024). In Indonesia, deforestation has shown a significant declining trend from 2015 to 2022, although fluctuations may still occur in the future (Qomaria, 2024). Based on statistics from the Indonesian Central Bureau of Statistics (BPS), the rate of deforestation has continued to decrease substantially between 2014–2015 and 2021–2022 (Badan Pusat Statistik, 2023). The highest deforestation occurred in 2014–2015, reaching 1,092,181.5 hectares, followed by fluctuations in 2018–2019 and 2020–2021. Consequently, effective regulations for forest management based on forest types are needed not only to support community welfare but also to preserve forest sustainability at the national level.

To support data-driven decision-making in sustainable forest management, analytical approaches that can identify structural patterns among regions are essential. Cluster analysis is one such approach widely used in data mining for automatically classifying objects into groups with similar characteristics (Hendrastuty, 2024). K-Means is a commonly applied clustering algorithm that partitions data into k clusters by minimizing the distance between objects

[*] Corresponding author.
E-mail address: sittimasyitahmr@unm.ac.id

and their corresponding cluster centers (Pratama et al., 2023). Meanwhile, the K-Medoids algorithm uses actual data points as cluster representatives (Srikandi & Yurinanda, 2025) and is known for being more robust to outliers than K-Means (Sukmayadi et al., 2021). Although both algorithms share similar conceptual foundations, they differ in the way cluster centers are determined.

Given these differences, comparing the performance of K-Means and K-Medoids offers a compelling analytical challenge. Previous studies have conducted similar comparisons. For example, Ilmi et al. (2024) found that while K-Means produced more varied cluster formations for hotspot data in Kalimantan, K-Medoids resulted in better average SSE values. Similarly, Khoirunnisa and Rahmawati (2024) reported that K-Means resulted in a lower Davies–Bouldin Index (0.425) than K-Medoids (0.939) when clustering natural disaster intensity in Indonesia. These findings highlight the importance of selecting an appropriate clustering algorithm depending on the dataset characteristics. Therefore, the present study aims to compare the performance of K-Means and K-Medoids in clustering Indonesian provinces based on the extent of forest areas by type.

2. Literature Review

2.1. Optimal Cluster Determination

In this study, the optimal number of clusters was selected using the Davies–Bouldin Index (DBI). DBI is a metric used to evaluate the quality of clustering in data analysis. Its purpose is to assess how well a clustering algorithm separates different groups of data while maintaining compactness within each cluster. A lower DBI value indicates better clustering performance. The formula used to calculate DBI is presented in Equation 1.

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{i\neq j}\frac{(a_1+a_j)}{(d(c_i,c_j))} \tag{1}$$

where n is the number of clusters, a_i and a_j represent the average distance of all members within clusters i and j, respectively, and $d(c_i, c_j)$ denotes the distance between the centroids of the two clusters.

2.2. K-Means Clustering

K-Means Clustering is a clustering method that utilizes the distance between objects, where the resulting distances reflect the degree of similarity between them. This method is a non-hierarchical clustering technique and is advantageous due to its ability to group large datasets quickly and efficiently. In K-Means, data are divided into several groups in which each cluster contains objects that are similar to one another but distinct from objects in other clusters. The objective is to minimize the variation within clusters and maximize the variation between clusters. The K-Means clustering algorithm follows the procedure outlined below.

- Determine the optimal number of clusters.
- Initialize the centroids randomly according to the predefined number of clusters.
- Calculate the distance between each data point and every centroid using Equation (2):

  $$d(x_i,\mu_j) = \sqrt{\sum(x_i-\mu_j)^2} \tag{2}$$

  where x_i represents a data point and μ_j denotes the centroid of cluster j.
- Assign each data point to the nearest centroid based on the smallest distance.
- Update the centroid values using the new mean of all points assigned to each cluster, as shown in Equation (3):

  $$\mu_j(t+1) = \frac{1}{N_{sj}}\sum_{j\epsilon sj} x_j \tag{3}$$

  where $\mu_j(t+1)$ is the updated centroid for iteration (t+1), and N_{sj} is the number of data points in cluster sj.
- Repeat the process until the centroid values no longer change or the algorithm reaches the predetermined maximum number of iterations.

2.3. K-Medoids Clustering

K-Medoids Clustering, also known as the Partitioning Around Medoids (PAM) algorithm, is a partition-based clustering method used to group a set of n objects into a specified number of clusters. Unlike K-Means, which uses the mean of objects as the cluster center, K-Medoids selects actual data points (medoids) as the center of each cluster. The K-Medoids algorithm follows the procedural steps outlined below.

- Initialize the number of clusters.
- Calculate the Euclidean distance to assign each data object to the nearest cluster using Equation (4):

$$d_{ij} = \sqrt{\sum_{a=1}^{p}(X_{ia} - X_{ja})^2} = \sqrt{(X_i - X_j)^I (X_i - X_j)} \quad (4)$$

  where p represents the number of variables, X is the covariance matrix, and i, j, n denote integer indices of the data objects.
- Select random candidates for new medoids from each group of items.
- Reassign each data point to the nearest medoid candidate to update the cluster composition.
- Compute the total deviation (S) by comparing the total distance of the previous medoid configuration with that of the new configuration. If S=0, replace the previous medoid with the new candidate in order to form the updated set of K-medoids.
- Repeat Steps 3–5 until no further changes occur in the medoid positions, indicating that the cluster centers and their respective cluster memberships have stabilized.

2.4. Cluster Quality Evaluation

To evaluate the quality of the clusters formed in this study, the Silhouette Coefficient method was employed. This method integrates two core concepts in cluster validation: cohesion and separation. Cohesion measures the average proximity of an object to other objects within the same cluster, while separation measures the average distance of that object to the nearest neighboring cluster (Rousseeuw, 1987). The distance between observations was computed using the Euclidean distance metric.

The silhouette value for each observation is calculated based on the balance between cohesion and separation. Meanwhile, the overall silhouette score for a clustering solution with k clusters is defined as the average silhouette value across all observations within those clusters. The silhouette coefficient can be expressed using Equation (5) as follows (Kaufman & Rousseeuw, 2005):

$$sil(c) = sil(k)\frac{1}{|k|}\sum_{i=1}^{k} sil(c_i) \quad (5)$$

where sil(k) denotes the overall silhouette value of the clustering solution, $|k|$ is the number of clusters, and sil(ci) is the average silhouette value for cluster i.

The Silhouette Coefficient ranges from 0 to 1, where higher values indicate better-defined and more coherent clusters. The interpretation categories for the Silhouette Coefficient are presented in Table 2 (Kodinariya & Makwana, 2013).

Table 1. Interpretation of Silhouette Coefficient Values

Silhouette Coefficient	**Interpretation**
0.7 < SC ≤ 1	Strong cluster structure
0.5 < SC ≤ 0.7	Moderate cluster structure
0.25 < SC ≤ 0.5	Weak cluster structure
SC ≤ 0.25	No apparent structure

3. Research Method

3.1. Data Source

The data used in this study are secondary data obtained from the official website of Statistics Indonesia (BPS RI) (www.bps.go.id), specifically the dataset titled *Forest Area and Aquatic Conservation Area by Province and Forest Function, 2023*. Only variables relevant to the objectives of this research were selected for analysis. The dataset used in this study is presented in Table 2.

To examine the characteristics of the data, descriptive statistical analysis was performed using IBM SPSS Statistics. Subsequently, the clustering analysis was conducted using the R-Studio software environment.

Table 2. Research Dataset

Province	Protected Forest (Thousand Ha)	Nature Reserves & Wildlife Conservation (Thousand Ha)	Limited Production Forest (Thousand Ha)	Permanent Production Forest (Thousand Ha)	Convertible Production Forest (Thousand Ha)
Aceh	1781678	1058364	145178,3	549794,9	15374,69
North Sumatera	1206881	427008	641769	704452	75684
West Sumatera	791671	806939	233211	360608	187629
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
Riau	233910	630753	1017318	2339578	1185433
Jambi	179588	685471	258285	963792	11399
North Maluku	584058	218499	666851	481730	564082
West Papua	1631589	2640258	1778480	2188160	1474650
Papua	7815283	7755284	5961240	4739327	4116365

For the province of North Kalimantan, the data are still aggregated under East Kalimantan; therefore, East Kalimantan and North Kalimantan are treated as a single integrated region in this study.

3.2. Research Method

This study uses secondary data from Statistics Indonesia (BPS) regarding forest area by province and forest function in 2023. East Kalimantan and North Kalimantan are treated as a single region because the data for North Kalimantan remain aggregated under East Kalimantan. Five variables were analysed: Protected Forest, Nature and Wildlife Conservation, Limited Production Forest, Permanent Production Forest, and Convertible Production Forest.

Descriptive statistics were generated using IBM SPSS Statistics to examine data characteristics. Prior to clustering, all variables were normalized using the Min–Max method to ensure comparability across scales.

Two clustering approaches were applied: K-Means Clustering and K-Medoids Clustering. K-Means groups data by minimizing within-cluster variance, while K-Medoids selects actual data points as medoids and is more robust to outliers. Both methods use Euclidean distance for assigning data to clusters.

The optimal number of clusters was determined using the Davies–Bouldin Index (DBI), where lower values indicate better cluster compactness and separation. Cluster performance was evaluated using the Silhouette Coefficient, which measures cohesion and separation, with values ranging from 0 (poor) to 1 (excellent). All clustering analyses were performed using R-Studio.

4. Results and Discussions

4.1. Descriptive Analysis

To understand the characteristics of the data used in this study, a descriptive statistical analysis was conducted. The results are presented in Table 3.

Based on Table 3, all 33 provinces included in the analysis have protected forest areas, with an average of 894,574.64 hectares. However, the size of protected forests varies substantially across provinces, ranging from 44.76 hectares to 7,815,283 hectares. This wide variation contributes to a standard deviation that exceeds the mean value.

The results also show that some provinces do not have Limited Production Forest or Convertible Production Forest areas, indicated by the minimum value of 0 for these variables. Meanwhile, other provinces possess both forest types, although their sizes also vary considerably. Overall, the data exhibit high variability across all forest functions, reflecting substantial differences in forest resource distribution among Indonesian provinces.

Table 3. Descriptive Statistics of Research Variables

Variable	N	Minimum	Maximum	Mean	Std. Deviation
Protected Forest	33	44.76	7,815,283	894,574.64	1,422,025.72
Nature and Wildlife Conservation	33	910.34	7,755,284	830,673.48	1,393,931.06
Limited Production Forest	33	0.00	5,961,240	812,084.86	1,411,447.33
Permanent Production Forest	33	158.35	4,739,327	883,975.28	1,251,312.56
Convertible Production Forest	33	0.00	4,116,365	386,708.42	872,432.38
Valid N (listwise)	33				

4.2. Model Development

4.2.1. K-Means Clustering

The clustering analysis using the K-Means method began with determining the optimal number of clusters. In this study, the selection of the appropriate number of clusters was based on the Davies–Bouldin Index (DBI). The DBI values were computed for cluster numbers ranging from *k = 2* to *k = 6*. The results are presented in Table 4.

Table 4. Davies–Bouldin Index for Different Numbers of Clusters

Number of Clusters	DBI Value
2	0.708
3	0.558
4	0.515
5	0.764
6	0.698

As shown in Table 4, the lowest DBI value is obtained when k = 4, with a value of 0.515. Since a lower DBI indicates better-defined clusters, those that are more compact and well separated k = 4 was identified as the optimal number of clusters for the K-Means clustering analysis.

4.2.2. K-Medoid Clustering

Similar to the K-Means procedure, the K-Medoids clustering method also requires determining the optimal number of clusters. In this study, the selection of the optimal cluster number was based on the Davies–Bouldin Index (DBI). The DBI values for cluster numbers ranging from *k = 2* to *k = 6* are presented in Table 5.

Table 5. Davies–Bouldin Index for Different Numbers of Clusters in K-Medoids

Number of Clusters	DBI Value
2	0.806
3	0.559
4	0.834
5	0.752
6	0.685

Based on Table 5, the lowest DBI value is obtained when k = 3, with a value of 0.559. Since a lower DBI indicates clusters that are more compact and better separated, it can be concluded that k = 3 represents the optimal number of clusters for the K-Medoids clustering analysis.

4.3. Evaluation of the Best Model

A comparison was then conducted between the clustering results obtained using the K-Means method with *k = 4* and the K-Medoids method with *k = 3*. The evaluation of both models was performed using the Silhouette Coefficient, which measures the compactness and separation of clusters. The evaluation results are presented in Table 6.

Table 6. Cluster Evaluation Results

Method	Number of Clusters	Silhouette Coefficient
K-Means Clustering	4	0.610
K-Medoids Clustering	3	0.644

Based on Table 6, the K-Medoids Clustering method yields a higher Silhouette Coefficient compared to the K-Means method. This indicates that the K-Medoids algorithm produces more well defined and better-separated clusters. Therefore, K-Medoids Clustering is identified as the superior method for grouping Indonesian provinces based on the area of forest land categorized by forest type.

4.4. Visualization of Clustering Analysis

The clustering analysis using the K-Medoids algorithm with *k = 3* was then performed with the assistance of the R programming language. The resulting clusters were visualized to provide a clearer depiction of the distribution pattern of each group based on the characteristics used in the analysis. The cluster distribution is illustrated using a scatterplot, as presented in Figure 1.

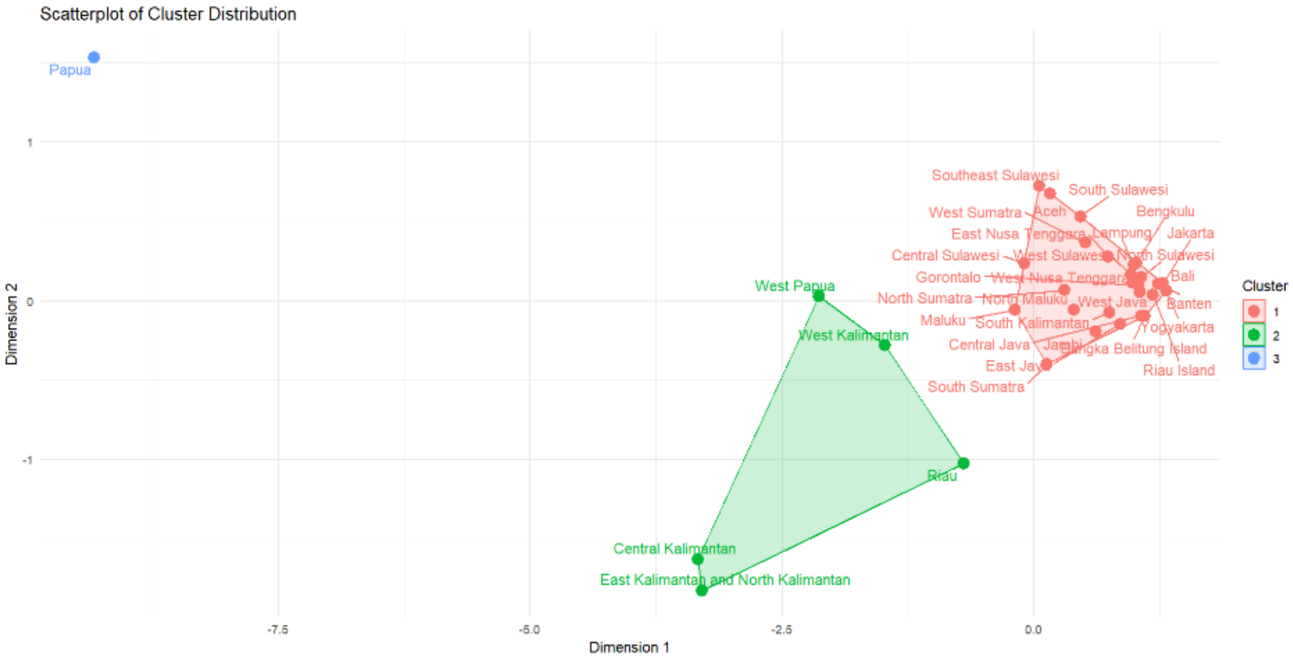

Figure 1. Cluster Distribution Using the K-Medoids Clustering Method

Based on Figure 1, three distinct clusters can be identified. **Cluster 1**, shown in red, represents the largest group in the dataset. **Cluster 2**, shown in green, forms a medium-sized cluster. Meanwhile, **Cluster 3** appears as an outlier cluster, indicating that the provinces within this group exhibit characteristics that differ substantially from the rest of the data. Each cluster demonstrates unique features that distinguish it from the others. The characteristics of each cluster along with their respective members are presented in Table 7.

The clustering results can also be visualized using a thematic map, as shown in Figure 2. Based on the map, Cluster 1 is distributed across most provinces in Sumatra, all provinces on Java extending to Bali and the Nusa Tenggara Islands, the entire Sulawesi region, the Maluku Islands, and a small part of Kalimantan. Cluster 2 is dominated by most provinces in Kalimantan, along with one province in Sumatra and one province in Papua. Meanwhile, Cluster 3 contains only a single province is Papua which stands as the sole member of this cluster.

Table 7. Cluster Characteristics and Members

Cluster	Characteristics	Cluster Members
Cluster 1	This cluster represents provinces characterized by relatively dominant areas of protected forests (*hutan lindung*) and nature conservation forests compared to production forests, including both permanent and limited production forests. It also includes provinces with medium to small forest areas across all forest categories.	Aceh, North Sumatra, West Sumatra, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, South Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku.
Cluster 2	This cluster represents provinces with very large and dominant areas of permanent production forests, limited production forests, and convertible production forests.	East Kalimantan and North Kalimantan (combined region), Central Kalimantan, West Kalimantan, West Papua, Riau.
Cluster 3	This cluster consists solely of the province of Papua, which has exceptionally large forest areas across all forest types. The total forest area in Papua surpasses all other provinces, making this cluster a natural outlier due to values that are significantly higher than the others.	Papua.

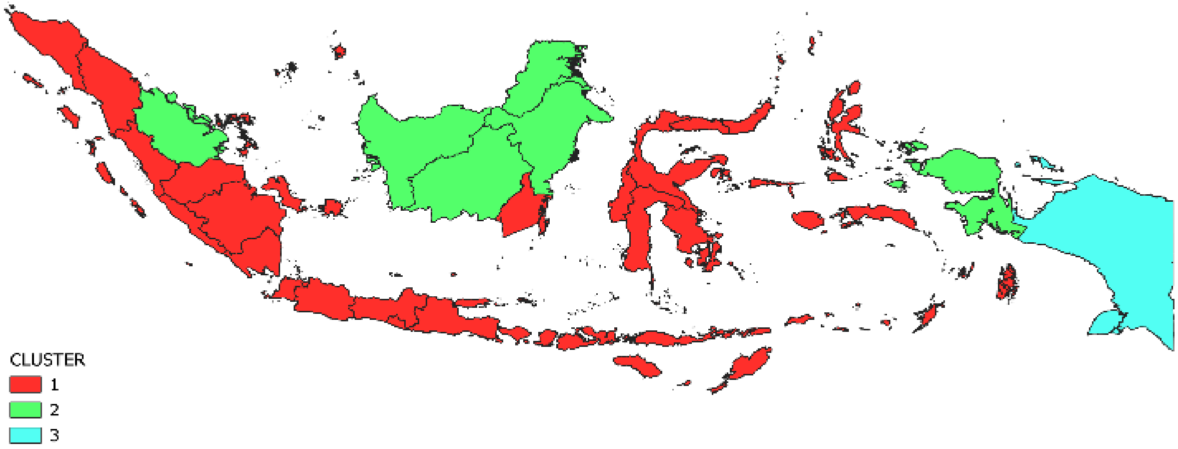

Figure 2. Thematic Map of Clusters

5. Conclusion

The clustering analysis of forest area by type using the K-Means and K-Medoids algorithms produced distinct outcomes. In the K-Means Clustering analysis, the optimal number of clusters—determined using the lowest DBI value—was four clusters, with a Silhouette Coefficient of 0.610. This result also showed that two of the clusters contained only a single member. Meanwhile, in the K-Medoids Clustering analysis, the optimal number of clusters based on the lowest DBI value was three clusters, with a higher Silhouette Coefficient of 0.644, and only one cluster consisted of a single member.

Based on the overall cluster quality, the K-Medoids Clustering method with three clusters demonstrates superior performance in grouping Indonesian provinces according to the area of forest types.

References

Akram, A., Risal, N., Maryani, D., Fadillah, N., Alviadi, A., & Risal, N. (2024). Implementasi K-Means clustering untuk rekomendasi kelas unggulan di SMK 1 Teknologi dan Rekayasa Mimika. *JESSI Journal of Embedded Systems, Security & Intelligent Systems*, 5(3), 255–261.

Badan Pusat Statistik. (2024). Angka deforestasi (netto) Indonesia di dalam dan di luar kawasan hutan tahun 2013-2022 (Ha/Th). Badan Pusat Statistik. Retrieved from https://www.bps.go.id/id/statistics-table/1/MjA4MSMx/angka-deforestasi--netto--indonesia-di-dalam-dan-di-luar-kawasan-hutan-tahun-2013-2022--ha-th-.html

Bakri, R., Sobirov, B., Astuti, N. P., Ahmar, A. S., & Singh, P. K. (2025). A new framework for dynamic educational marketing segmentation in student recruitment: Optimizing fuzzy C-Means with metaheuristic techniques. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 9(3), 659–669. https://doi.org/10.29207/resti.v9i3.6515

Dinh, D. T., Fujinami, T., & Huynh, V. N. (2019). Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. *Communications in Computer and Information Science*, 1103, 1–17. https://doi.org/10.1007/978-981-15-1209-4_1

Fatkhudin, A., Khambali, A., Artanto, F. A., Putra Zade, N. A., & Muhammadiyah Pekajangan Pekalongan. (2023). Implementasi algoritma clustering K-Means dalam pengelompokan mahasiswa studi kasus (Prodi Manajemen Informatika). *Jurnal Minfo Polgan*, 12(2), 777–783. https://doi.org/10.33395/jmp.v12i2.12494

Faturrahman, S., Hariani, S., & HariniKusumawati, R. (2023). Evaluasi clustering K-Means dan K-Medoid pada persebaran Covid-19 di Indonesia dengan metode Davies-Bouldin Index (DBI). *Jurnal Mnemon*, 6(2), 117–128.

Hafid, H., Meliyana, S. M., Muthahharah, I., & Mar'ah, Z. (2025). Implementation K-Medoids algorithm for clustering Indonesian provinces by poverty and economic indicators. *Quantitative Economics and Management Studies*, 6(2), 219-225. https://doi.org/10.35877/454RI.qems3940

Hafid, H., & Meliyana, S. M. (2024). Implementation of K-Median algorithm for the regencies clustering in South Sulawesi Province based on food commodity yields. *Journal of Mathematics, Computation and Statistics*, 7(2), 283-294. https://doi.org/10.35580/jmathcos.v7i2.3674

Ilmi, H. M., Kurniawan, M., Al Faruq, U., & Muhima, R. R. (2024). Comparison of K-Means and K-Medoids for hotspot data clustering on the island of Kalimantan. *Jurnal SimanteC*, 13(1), 33–40.

Karo, I. M. K., Dewi, S., Mardiana, F., Ramadhani, F., & Harliana, P. (2023). K-means and K-medoids algorithm comparison for clustering forest fire location in Indonesia. *Jurnal ECTIPE Elektronika, Control, Telecommunication, Information, and Power Engineering*, 10(1), 86–94. https://doi.org/10.33019/jurnalecotipe.v10i1.3896

Noer, M., & Dimyati, M. (2024). Systematic literature review: Pola spasial, tren dan dinamika deforestasi hutan dalam perspektif penginderaan jauh. *Geografi: Jurnal Kajian, Penelitian, dan Pengembangan Pendidik*, 12(1), 412–423.

Pratama, A. R., Maulana, B., Rianda, R. D., & El Hasyim, S. (2023). Perbandingan algoritma K-Means dan K-Medoids untuk pengelompokan data penjualan video game di Amerika Utara. *Indonesian Journal of Information Research & Software Engineering*, 3(2), 111–118.

Pratiwi, A. S., Syartinilia, & Pravitasari, A. E. (2024). Perubahan tutupan lahan, degradasi, dan deforestasi hutan di Kabupaten Nabire periode 2000-2019. *Jurnal Lanskap Indonesia*, 16(2), 199–207. https://doi.org/10.29244/jli.v16i2.54249

Qomaria, R. (2024). Pengelompokan kasus deforestasi di Indonesia menggunakan metode K-Means. *B.S. thesis*, Department of Mathematics, UIN Sunan Ampel Surabaya, Surabaya, Indonesia.

Shafitri, L. D., Prasetyo, Y., & Hani'ah. (2018). Analisis deforestasi hutan di Provinsi Riau dengan metode polarimetrik dalam pengindraan jauh. *Jurnal Geodesi Undip*, 7, 1–11.

Sukmayadi, C., Primajaya, A., & Maulana, I. (2021). Penerapan algoritma K-Medoids dalam menentukan daerah rawan banjir di Kabupaten Karawang. *INFORMAL Informatics Journal*, 6(3), 187. https://doi.org/10.19184/isj.v6i3.25423

Srikandi, & Yurinanda, S. (2025). Analisis cluster program anggaran untuk meningkatkan efisiensi dengan metode K-Medoids di Sekretariat DPRD Provinsi Jambi. *STATMAT (Jurnal Statistik dan Matematika)*, 7(2), 180–200.