ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

COMPARATIVE ANALYSIS OF K-MEANS, K-MEDOIDS, AND DBSCAN FOR SPATIAL CLUSTERING OF POSITIVE CLOUD-TO-GROUND LIGHTNING IN TANGERANG CITY

Adi Pranata Ranta, Marzuki Sinambela✉, Agustina Rachmawardani, Deni Septiadi
Sekolah Tinggi Meteorologi Klimatologi dan Geofisika, Tangerang, Indonesia
Email: sinambela.m@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol10No2.pp451-462

ABSTRACT

Cloud-to-Ground (CG) lightning is one of the most destructive atmospheric phenomena due to its potential to damage electrical infrastructure, communication networks, and public facilities. Understanding its spatial distribution is essential for disaster mitigation, infrastructure protection, and urban planning. This study compares the performance of the K-Means, K-Medoids, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms in identifying the spatial distribution of Positive Cloud-to-Ground (CG+) lightning strikes in Tangerang City, Indonesia. The analysis utilizes CG+ lightning observations collected by the Tangerang Geophysical Station of the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG) during 2025. Prior to clustering, the dataset underwent data cleaning, CG+ event selection, administrative boundary filtering, feature selection, coordinate transformation from WGS84 to UTM, and Min–Max normalization. The clustering performance of each algorithm was assessed using the Silhouette Coefficient and Davies–Bouldin Index (DBI). The evaluation results indicate that both K-Means and K-Medoids achieved the best performance with K = 3. K-Means obtained the highest Silhouette Coefficient (0.445640) and the lowest DBI (0.789620), while K-Medoids produced comparable results with a Silhouette Coefficient of 0.444728 and a DBI of 0.794027. For DBSCAN, the optimal parameter combination (ε = 200 m and MinPts = 6) generated 64 clusters and 170 noise observations, yielding a Silhouette Coefficient of 0.059880 and a DBI of 1.452474. These findings demonstrate that partition-based clustering approaches provide a more reliable representation of the spatial distribution of CG+ lightning in Tangerang City than the density-based DBSCAN approach.

Keyword: Cloud-to-Ground Lightning, DBSCAN, K-Means, K-Medoids, Spatial Clustering.

INTRODUCTION

Cloud-to-Ground (CG) lightning is recognized as one of the most hazardous atmospheric phenomena because it transfers electrical energy directly from thunderclouds to the Earth's surface. These discharges may cause substantial damage to electric power systems, telecommunication facilities, transportation infrastructure, and other public assets, while also posing risks to human safety. Previous studies have shown that lightning occurrence is influenced by atmospheric instability, climatic conditions, and land-surface characteristics, resulting in considerable spatial variability across different regions (Shi et al., 2025).

CG lightning is commonly classified according to the polarity of the electric charge into Positive Cloud-to-Ground (CG+) and Negative Cloud-to-Ground (CG−) lightning. Although CG− events occur more frequently, CG+ lightning generally produces higher peak currents, longer continuing currents, and greater discharge energy. Consequently, CG+ strikes are more likely to damage critical infrastructure such as transmission lines, substations, communication towers, and other electrical facilities (Jiang et al., 2023). Understanding the spatial distribution of CG+ lightning is therefore essential for disaster risk reduction, infrastructure planning, and urban resilience.

The spatial distribution of lightning is rarely uniform because it is affected by a combination of meteorological and environmental factors, including convective activity, terrain characteristics, land use, and urban development. Urbanization may further modify local atmospheric conditions through the urban heat island effect, changes in surface roughness, and increased aerosol concentrations, all of which can influence lightning occurrence (Wang et al., 2020). Identifying areas with high lightning activity is therefore important for supporting mitigation strategies and improving infrastructure protection.

Recent developments in geospatial data analysis have enabled clustering techniques to be widely applied for identifying spatial patterns in environmental datasets. Among the most frequently used clustering methods are K-Means, K-Medoids, and Density-Based Spatial Clustering of Applications with Noise

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

(DBSCAN). K-Means partitions observations by minimizing the distance between data points and cluster centroids, whereas K-Medoids represents each cluster using an actual observation, making it less sensitive to outliers. In contrast, DBSCAN groups observations according to local point density and can detect clusters with arbitrary shapes while simultaneously identifying noise (Agresto et al., 2024; Schubert & Rousseeuw, 2021).

Several studies have applied clustering algorithms to spatial datasets, including flood mapping and environmental analysis (Ashari et al., 2023; Nugrahanto & Yunin Sofro, 2025; Salman & Fauziah, 2023). However, comparative studies focusing on the spatial distribution of Positive Cloud-to-Ground (CG+) lightning remain limited, particularly in rapidly developing urban regions such as Tangerang City. Moreover, relatively few studies have evaluated K-Means, K-Medoids, and DBSCAN using the same lightning dataset and identical validation metrics, making it difficult to objectively determine the most appropriate algorithm for this application.

This study presents an in-depth comparison of the K-Means, K-Medoids, and DBSCAN algorithms using the same Positive Cloud-to-Ground (CG+) lightning dataset, identical preprocessing techniques, and consistent internal validation metrics—unlike previous studies, which mostly applied clustering algorithms in isolation or focused on general spatial clustering problems. The strengths and weaknesses of partition-based and density-based clustering techniques for the spatial analysis of CG+ lightning can be objectively evaluated thanks to this integrated experimental framework. The results of this study also provide useful recommendations for selecting appropriate clustering techniques to improve infrastructure protection, disaster mitigation, and lightning hazard assessment in urban environments in the future.

This study evaluates how well the K -Means, K-Medoids, and DBSCAN methods cluster the spatial distribution of Positive Cloud-to-Ground (CG+) lightning observations recorded by the Tangerang Geophysical Station of the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG) in 2025, in order to validate the proposed comparative framework. Normalized UTM coordinates were used in this study, while the Davies-Bouldin Index and the Silhouette Coefficient were used to evaluate the quality of the clustering. It is hoped that these findings will determine which clustering approach best captures the spatial characteristics of CG+ lightning and serve as a reproducible guideline for further spatial lightning research.

RESEARCH METHOD

Data Collection

The dataset used in this study consists of Cloud-to-Ground (CG) lightning observations provided by the Tangerang Geophysical Station of the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG). This dataset covers the observation period from January through December 2025 and includes lightning events recorded in and around the city of Tangerang. Each observation includes geographic coordinates and lightning polarity information, enabling a spatial analysis of lightning events(Jiang et al., 2023; Paski et al., 2017).

Observations of positive (CG+) and negative (CG−) cloud-to-ground lightning were initially included in the dataset. Only CG+ data were selected for further analysis because the objective of this study was to determine the spatial distribution of positive lightning events. Positive cloud-to-ground lightning is more critical for disaster mitigation and infrastructure protection because it is known to produce higher peak currents and more damaging effects than negative lightning(Jiang et al., 2023).

The raw data set includes several attributes that describe each lightning strike. However, only the variables necessary for spatial clustering were retained: ID, Longitude, and Latitude. Longitude and latitude indicate the geographic location of each lightning strike, while the ID attribute serves as a unique identifier for each observation.

Table 1. Dataset Description

Item	Description
Data Source	Tangerang Geophysics Station – BMKG
Observation Period	January – December 2025
Lightning Type	Cloud-to-Ground (CG)
Selected Data	Positive Cloud-to-Ground (CG+)
Coordinate System	WGS84
Selected Features	ID, Longitude, Latitude

Data Preprocessing

Preliminary data processing is performed to improve the quality and consistency of the lightning dataset before applying the clustering algorithm. Since clustering methods are highly influenced by data quality and spatial representation, preliminary processing is a crucial step in ensuring reliable clustering results. The preprocessing workflow consists

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

of six sequential stages: data cleaning, data selection, administrative boundary filtering, feature selection, coordinate transformation, and data normalization.

1) Data Cleaning

The first preprocessing step involves removing duplicate records, incomplete observations, and invalid geographic coordinates from the dataset. Records that do not include longitude or latitude values are excluded because accurate spatial coordinates are required for distance-based clustering. This process ensures that only valid lightning observations are used during the analysis(Ashari et al., 2023; Han et al., 2012).

2) Data Selection

The dataset initially contained both CG+ and CG− lightning events. To meet the objective of this study, only observations classified as Positive Cloud-to-Ground (CG+) were retained. All negative cloud-to-ground lightning observations were removed before further processing(Jiang et al., 2023)

3) Administrative Boundary Filtering

After the data selection process, CG+ lightning observations were spatially filtered using the official administrative boundaries of Tangerang City. Lightning events located outside the study area were excluded to ensure that the clustering results reflected only the spatial distribution of lightning within the city.

4) Feature Selection

Only three attributes were selected for the clustering analysis: ID, Longitude, and Latitude. The ID attribute was retained as a unique identifier, while longitude and latitude were used as spatial variables to identify lightning strike clusters(Ashari et al., 2023).

5) Coordinate Transformation

Lightning strike coordinates were initially entered using geographic coordinates (latitude and longitude) in the World Geodetic System 1984 (WGS84). These coordinates were converted to the Universal Transverse Mercator (UTM) coordinate system because the clustering technique relies on Euclidean distance calculations. By converting geographic coordinates into planar coordinates expressed in meters, this transformation improves the accuracy of spatial distance calculations(Karney, 2012).

6) Data Normalization

Finally, the transformed UTM coordinates were normalized using the Min–Max normalization method. Normalization rescales each feature to the range of 0–1, preventing variables with larger numerical values from dominating the distance calculation during clustering. The Min–Max normalization is defined as:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

Where**:**

- x= original value,
- $x_{\min}$= minimum value,
- $x_{\max}$= maximum value,
- x'= normalized value.

The normalized dataset was subsequently used as the input for the K-Means, K-Medoids, and DBSCAN clustering algorithms.

Clustering Methods

This study used three clustering algorithms—K-Means, K-Medoids, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)—to identify the spatial distribution of Positive Cloud-to-Ground (CG+) lightning strikes in Tangerang City. The entire clustering process was performed using normalized UTM coordinates obtained from the preprocessing stage. The clustering results generated by each algorithm were then evaluated using the Silhouette Coefficient and the Davies–Bouldin Index to determine the most appropriate algorithm for identifying lightning-prone areas (Kaufman & Rousseeuw, 1990).

To ensure a systematic and reproducible evaluation, the clustering parameters were determined experimentally. For the K-Means and K-Medoids algorithms, the number of clusters (K) varied from 2 to 10 to evaluate various spatial partitioning scenarios while avoiding clusters that were too small or too fragmented. For DBSCAN, the neighborhood radius (ε) was tested at 200 m, 300 m, 400 m, and 500 m to represent different neighborhood scales within the study area. The minimum number of neighboring points (MinPts) was varied between 4 and 6, following commonly applied recommendations for two-dimensional spatial datasets. The optimal parameter configuration for each algorithm was then selected based on the highest Silhouette Coefficient and the lowest Davies–Bouldin Index.

K-Means Clustering

Prior to the clustering process, both the K-Means and K-Medoids algorithms experimentally varied the number of clusters (K) from 2 to 10. This range was used to evaluate various levels of geographic partitioning without producing clusters that were too small or fragmented. A similar experimental range is often used in previous clustering studies, where internal validation metrics such as the Davies-Bouldin Index and the Silhouette Coefficient are used to determine the

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ideal number of clusters (Ashari et al., 2023; Kaufman & Rousseeuw, 1990).

K-Means was applied as a partition-based clustering algorithm to group lightning events based on their spatial similarity. This algorithm begins by initializing k cluster centers and assigning each observation to the nearest cluster center based on Euclidean distance. After all observations have been assigned, each cluster center is recalculated as the average position of all observations within that cluster. This iterative process continues until the cluster center positions no longer change or the maximum number of iterations is reached(Ashari et al., 2023). The Euclidean distance used in this study is expressed as:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - c_{jk})^2} \qquad (2)$$

Where**:**

- $d(x_i, c_j)$= Euclidean distance between observation $x_i$and centroid c_j,
- m= number of features,
- x_{ik}= value of feature $k$of observation i,
- c_{jk}= value of feature $k$of centroid j.

The number of clusters was determined experimentally based on the clustering performance obtained during the evaluation stage.

K-Medoids Clustering

By selecting actual data points as cluster representatives (medoids), the K-Medoids method is used to improve the clustering's robustness against outliers. K-Medoids seeks observations that minimize the total distance to all other observations within the same cluster, unlike K-Means, which determines centroids using the arithmetic mean(Schubert & Rousseeuw, 2021).

The Partitioning Around Medoids (PAM) technique is used in the clustering process. First, k observations are selected as potential medoids. Euclidean distance is used to assign each lightning event to the nearest medoid. Next, whenever such a swap reduces the total clustering cost, the algorithm iteratively swaps a medoid with a non-medoid data point. When there is no further reduction in the clustering cost, the iteration ends. The Euclidean distance is calculated as:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - c_{jk})^2} \qquad (3)$$

Where**:**

- $d(x_i, c_j)$= Euclidean distance between observation $x_i$and centroid c_j,
- m= number of features,
- x_{ik}= value of the k-th feature of observation i,
- c_{jk}= value of the k-th feature of centroid j.

DBSCAN Clustering

DBSCAN is used to identify clusters based on the density of neighboring observations without needing to specify the number of clusters in advance. This algorithm groups observations based on two parameters: the neighborhood radius (ε) and the minimum number of neighboring observations (MinPts)(Agresto et al., 2024; Ester et al., 1996).

This study evaluated the optimal clustering configuration for determining the spatial distribution of Positive Cloud-to-Ground (CG+) lightning events using various combinations of DBSCAN parameters. To examine how the neighborhood size affects cluster formation, the neighborhood radius (ε) was experimentally varied to 200 m, 300 m, 400 m, and 500 m. These values were chosen to reflect a range of neighborhood sizes, from local to broader spatial connections, and to account for the physical extent of the study area. Meanwhile, the MinPts parameter was set to 4, 5, and 6 in accordance with standard guidelines for two-dimensional spatial datasets and to evaluate DBSCAN's sensitivity to different density thresholds (Agresto et al., 2024; Ester et al., 1996).

In this study, several combinations of DBSCAN parameters were evaluated to determine the most appropriate clustering configuration for identifying the spatial distribution of Positive Cloud-to-Ground (CG+) lightning events. The neighborhood radius (ε) was set to 200 m, 300 m, 400 m, and 500 m. while the minimum number of neighboring points (MinPts) was varied between 4, 5, and 6. During the clustering process, a lightning observation was classified as a core point if it had at least MinPts neighboring observations within its ε-neighborhood. Observations that were directly density-reachable from a core point were assigned to the same cluster, whereas observations that did not satisfy the density requirement were identified as noise. The clustering results obtained from each parameter combination were subsequently compared using internal validation metrics to determine the optimal DBSCAN configuration to identify the spatial distribution patterns of Positive Cloud-to-Ground (CG+) lightning in Tangerang City. The Euclidean distance between observations is defined as:

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2} \quad (4)$$

Where:

- $d(x_i, x_j)$ = Euclidean distance between observations $x_i$and x_j.

The clustering results obtained from each parameter combination were compared to identify the optimal DBSCAN configuration for CG+ lightning data.

Cluster Evaluation

The Silhouette Coefficient (SC) and the Davies–Bouldin Index (DBI), two measures of internal validation, were used to compare the clustering performance of the K-Means, K-Medoids, and DBSCAN methods. These measures were chosen for unsupervised learning tasks because they can assess the quality of clustering without requiring predefined class labels(Davies & Bouldin, 1979; Rousseeuw, 1987).

Compared to neighboring clusters, the Silhouette Coefficient measures how similar an observation is to the members of the cluster to which it is assigned. A higher average Silhouette value indicates better cluster separation and higher density (Rousseeuw, 1987). The Silhouette Coefficient is calculated as:

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (5)$$

Where:

- $a(i)$= average distance between observation $i$and all observations within the same cluster;
- $b(i)$= minimum average distance between observation $i$and observations in the nearest neighboring cluster.

The Davies–Bouldin Index (DBI) evaluates the average similarity between each cluster and its most similar neighboring cluster by considering both cluster compactness and cluster separation (Davies & Bouldin, 1979). The DBI is calculated as:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{S_i+S_j}{M_{ij}}\right) \quad (6)$$

Where:

- S_i= intra-cluster distance of cluster i;
- M_{ij}= distance between clusters $i$and j.

In this study, the algorithm with the highest Silhouette Coefficient and the lowest Davies–Bouldin Index was considered to provide the best clustering performance for identifying CG+ lightning in Tangerang City.

Using the same preprocessing techniques and internal validation metrics, the selected parameter ranges were systematically evaluated. To ensure that the parameter selection procedure could be replicated in future studies using comparable geographic datasets, the optimal parameter combinations for each clustering method were identified based on the highest Silhouette Coefficient and the lowest Davies–Bouldin Index.

RESULTS AND DISCUSSION

This section presents the experimental results obtained from the application of three clustering algorithms—K-Means, K-Medoids, and DBSCAN—to analyze the spatial distribution of Positive Cloud-to-Ground (CG+) lightning strikes in Tangerang City. This discussion follows the research workflow presented in the methodology, beginning with the preprocessing results, followed by the clustering results from each algorithm, and concluding with a comparative evaluation using the Silhouette Coefficient and the Davies–Bouldin Index.

Data Preprocessing Results

The preprocessing stage was conducted to ensure that the lightning dataset used in this study was complete, consistent, and suitable for clustering analysis. The original dataset consisted of 4,655,183 cloud-to-ground lightning observations recorded by the Tangerang Geophysical Station during the 2025 observation period. Data cleaning was performed first to remove duplicate records and observations containing incomplete or invalid coordinate information. After this process, 4,655,180 valid lightning observations remained.

Next, the dataset was filtered based on lightning polarity, selecting only Positive Cloud-to-Ground (CG+) lightning events. This filtering reduced the dataset to 1,356,780 observations. Since the objective of this study was to analyze the spatial distribution of CG+ lightning in Tangerang City, an additional spatial filtering process was applied using the official administrative boundaries of Tangerang City. As a result, the final dataset contains 10,235 CG+ lightning observations located within the study area.

Three attributes make up the selected dataset: ID, Latitude, and Longitude. To ensure accurate Euclidean distance calculations, the geographic coordinates are converted from the World Geodetic

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

System 1984 (WGS84) to the Universal Transverse Mercator (UTM) coordinate system. Finally, before the clustering process, the coordinate values are standardized using Min–Max normalization.

Table 2. Summary of Data Preprocessing

Preprocessing Stage	Number of Records
Original dataset	4,655,183
After data cleaning	4,655,180
Positive CG lightning (CG+)	1,356,780
CG+ within Tangerang City	10,235

The preprocessing stage successfully converted the original dataset into a structured spatial dataset suitable for clustering analysis. The final normalized dataset, containing 10,235 CG+ lightning observations, was then used as input for the K-Means, K-Medoids, and DBSCAN algorithms.

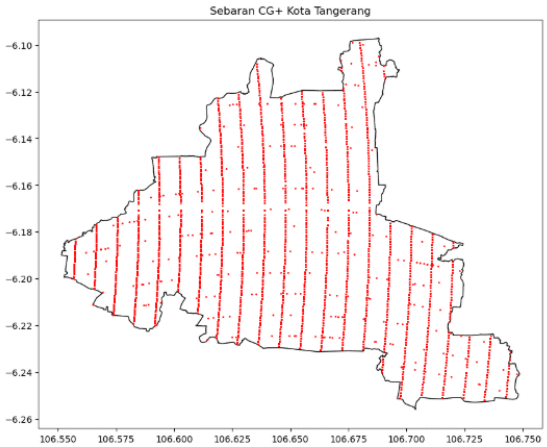


Figure 1. Spatial Distribution of Positive Cloud-to-Ground (CG+) Lightning Observations in Tangerang City after Preprocessing.

K-Means Clustering Results

The K-Means algorithm was applied to analyze the spatial distribution of positive Cloud-to-Ground (CG+) lightning observations in the Tangerang City area. Clustering was performed using normalized UTM coordinates obtained during the preprocessing stage. To determine the optimal number of clusters, several cluster configurations with K values ranging from K = 2 to K = 10 were evaluated using the Silhouette Coefficient and the Davies–Bouldin Index (DBI).

Table 3. Evaluation Results of K-Means Clustering

K	Silhouette Coefficient	Davies–Bouldin Index
2	0.415356	0.911188
3	**0.445640**	**0.789620**
4	0.411164	0.818292
5	0.401009	0.818395
6	0.408129	0.807048
7	0.390286	0.813631
8	0.391333	0.802972
9	0.388781	0.805953
10	0.379272	0.820153

As shown in Table 3, the quality of the clustering varies depending on the number of clusters. The highest Silhouette Coefficient value of 0.445640 and the lowest Davies–Bouldin Index value of 0.789620 were both obtained when K = 3. These results indicate that three clusters provide the most appropriate partitioning for the CG+ lightning dataset, offering the best balance between cluster compactness and inter-cluster separation.

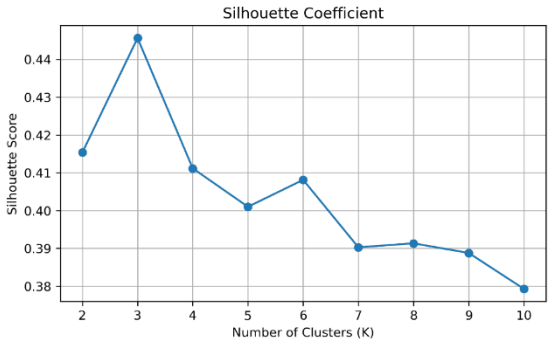


Figure 2. Silhouette Coefficient of K-Means for Different values of K.

As shown in Figure 2, the Silhouette Coefficient increases from 0.415356 at K = 2 to 0.445640 at K = 3, indicating an improvement in clustering quality. However, the coefficient gradually decreases as the number of clusters exceeds three, suggesting that adding more clusters reduces the cohesion and separation of the resulting groups. This trend confirms that K = 3 provides the optimal clustering configuration for the CG+ lightning dataset.

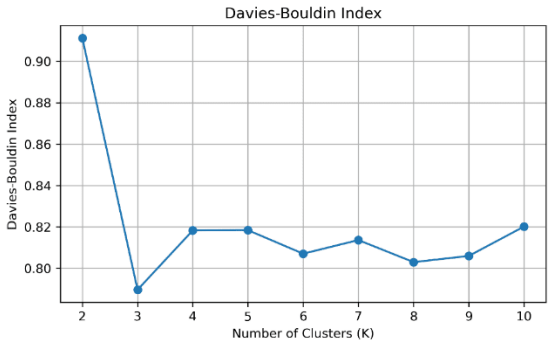


Figure 3. Davies–Bouldin Index of K-Means for Different values of K.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Figure 3 demonstrates that the DBI decreases significantly from 0.911188 at K = 2 to 0.789620 at K = 3, indicating better separation among clusters. For larger values of K, the DBI fluctuates between approximately 0.80 and 0.82, without showing further improvement. Therefore, the lowest DBI obtained at K = 3 confirms that this configuration produces the most compact and well-separated clusters.

Based on the results of this evaluation, the K-Means model with K = 3 was selected for further analysis. The spatial distribution of the resulting clusters is shown in Figure 4, while Figure 5 displays the same clustering results overlaid on an OpenStreetMap base map to facilitate geographical interpretation.

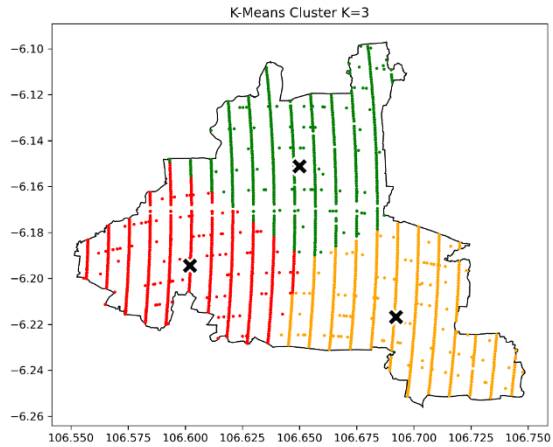


Figure 4. Spatial Distribution of CG+ lightning Clusters Generated using K-Means (K = 3).

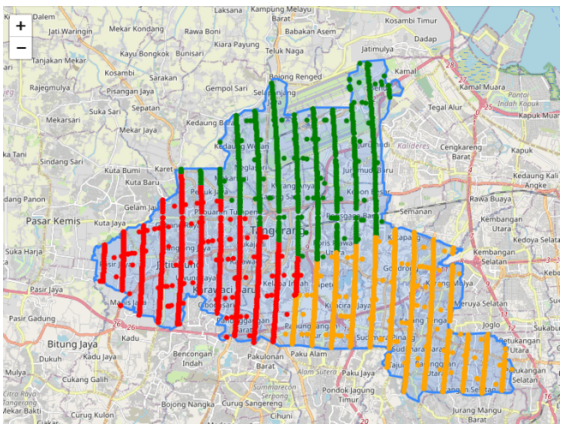


Figure 5. K-Means Clustering Results Overlaid on the OpenStreetMap Basemap.

Figures 4 and 5 show that the K-Means algorithm divides the CG+ lightning observations into three distinct spatial clusters. The western part of Tangerang City is dominated by Cluster 1 (red), while the northern and central areas are grouped into Cluster 0 (green). The eastern and southeastern regions are mostly classified as Cluster 2 (orange). The centroid of each cluster, marked with a black cross, is located near the geometric center of the respective cluster, indicating that the algorithm successfully identified representative locations for each spatial group.

This spatial distribution also shows that the cluster boundaries generally follow the geographic pattern of lightning observations, resulting in dense and clearly distinct clusters. Although the K-Means method partitions the data using Euclidean distance without considering administrative or physical boundaries, the resulting clusters show a clear regional separation that reflects the underlying spatial distribution of CG+ lightning events throughout the city of Tangerang.

K-Medoids Clustering Results

The K-Medoids algorithm was applied to analyze the spatial distribution of positive Cloud-to-Ground (CG+) lightning observations in Tangerang City. As in the K-Means experiment, the clustering process was performed using normalized UTM coordinates obtained during the preprocessing stage. To determine the optimal number of clusters, the number of clusters (K) was varied from 2 to 10, and each clustering result was evaluated using the Silhouette Coefficient and the Davies–Bouldin Index (DBI).

Table 4. Evaluation Results of K-Medoids Clustering

K	Silhouette Coefficient	Davies–Bouldin Index
2	0.417478	0.889170
3	**0.444728**	**0.794027**
4	0.410133	0.816253
5	0.400130	0.815423
6	0.403509	0.812688
7	0.390436	0.815270
8	0.391164	0.806353
9	0.371954	0.843266
10	0.374216	0.829191

The evaluation results show that the quality of clustering varies depending on the number of clusters selected. Among all the configurations tested, K = 3 achieved the highest Silhouette Coefficient (0.444728) and the lowest Davies–Bouldin Index (0.794027). These results indicate that dividing the dataset into three clusters provides the best balance between cluster compactness and cluster separation, making it the optimal clustering configuration for the K-Medoids algorithm.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

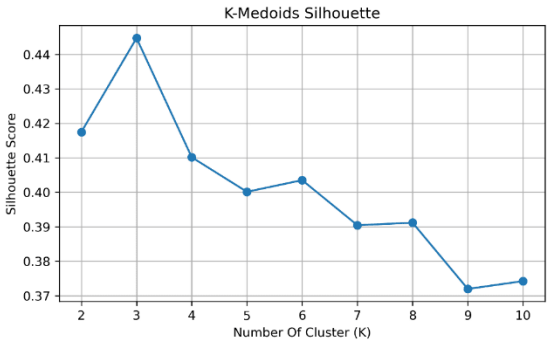


Figure 6. Silhouette Coefficient of K-Medoids for Different Values of K.

As shown in Figure 6, the Silhouette Coefficient increased from 0.417478 at K = 2 to 0.444728 at K = 3, indicating an improvement in clustering quality. However, the coefficient gradually decreases as the number of clusters exceeds three, indicating that adding more clusters reduces the cohesion and separation of the resulting groups. Therefore, the highest Silhouette value confirms that K = 3 provides the most appropriate clustering structure for the CG+ lightning dataset.

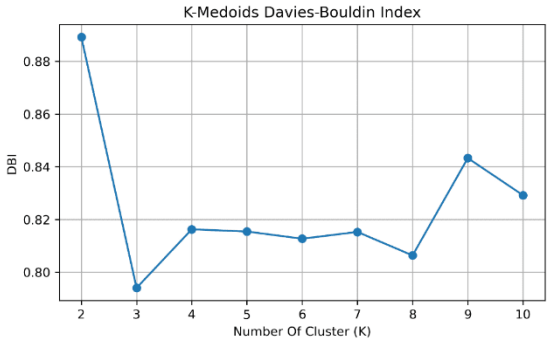


Figure 7. Davies–Bouldin Index of K-Medoids for Different Values of K.

Figure 7 shows that the DBI value decreases significantly from 0.889170 at K = 2 to 0.794027 at K = 3, indicating better inter-cluster separation. Although the DBI value remains relatively stable for larger values of K, none of the tested configurations produced a lower value than that obtained at K = 3. Therefore, both evaluation metrics consistently identify K = 3 as the optimal number of clusters.

Based on these results, the final K-Medoids model with K = 3 was selected for spatial analysis. The resulting clustering map is presented in Figure 9, while Figure 10 illustrates the same clustering result overlaid on the OpenStreetMap basemap to facilitate geographical interpretation.

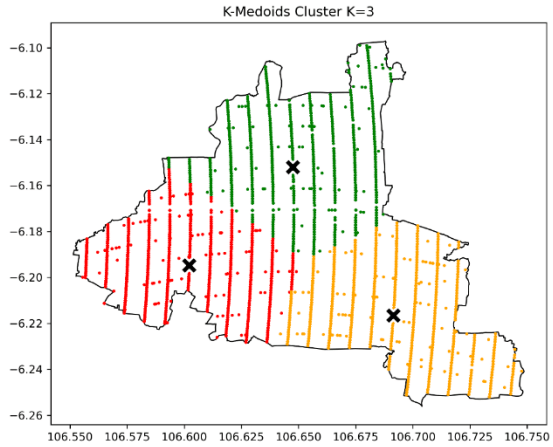


Figure 8. Spatial Distribution of CG+ Lightning Clusters Generated using K-Medoids (K = 3)

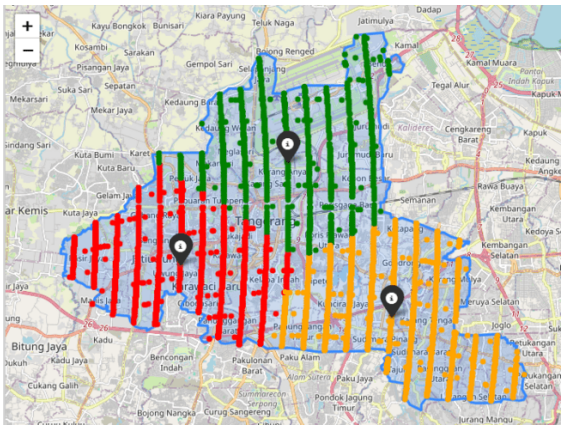


Figure 9. K-Medoids Clustering Results Overlaid on the OpenStreetMap Basemap.

Figures 8 and 9 show that the K-Medoids algorithm divides the CG+ lightning observations into three spatially distinct clusters. The western region of Tangerang City is primarily represented by Cluster 1 (red), the northern and central regions by Cluster 0 (green), and the eastern region by Cluster 2 (orange). Unlike K-Means, which calculates the cluster center (centroid) as the average position of each cluster, K-Medoids selects actual observations as representative centers (medoids). As a result, these representative points correspond to the actual lightning observations in the dataset.

In general, K-Medoids and K-Means produce fairly similar spatial distributions. This similarity indicates that the spatial distribution of CG+ lightning observations is quite stable, allowing both partition-based techniques to detect consistent clustering patterns throughout the city of Tangerang.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

DBSCAN Clustering Results

The DBSCAN algorithm is used to identify the spatial distribution of Positive Cloud-to-Ground (CG+) lightning observations based on local point density. Unlike K-Means and K-Medoids, DBSCAN does not require the number of clusters to be specified in advance. Instead, this algorithm determines clusters based on the neighborhood radius (ε) and the minimum number of neighboring observations (MinPts).

To evaluate the effect of parameter selection on clustering performance, four neighborhood radius values (ε = 200 m, 300 m, 400 m, and 500 m) were combined with three MinPts values (4, 5, and 6), resulting in a total of 12 clustering configurations. Each configuration was evaluated using the Silhouette Coefficient and the Davies–Bouldin Index (DBI).

Table 5. Evaluation Results of DBSCAN Clustering

ε (m)	MinPts	Number of Clusters	Noise	Silhouette	DBI
200	4	60	152	0.013395	1.573792
200	5	63	160	0.051927	1.468055
200	**6**	**64**	**170**	**0.059880**	**1.452474**
300	4	38	90	-0.036312	2.096057
300	5	38	92	-0.036242	2.095014
300	6	38	95	-0.036128	2.093585
400	4	23	25	-0.044648	2.454899
400	5	24	27	-0.043233	2.523615
400	6	24	27	-0.043233	2.523615
500	4	12	2	-0.093825	1.199603
500	5	12	2	-0.093825	1.199603
500	6	12	2	-0.093825	1.199603

The evaluation results show that the DBSCAN algorithm is highly sensitive to the choice of ε and MinPts values. As the neighborhood radius increases, the number of identified clusters gradually decreases, while the number of observations classified as noise also decreases. This behavior indicates that larger ε values tend to merge adjacent clusters into larger spatial groups.

Among all the parameter combinations tested, ε = 200 m and MinPts = 6 yielded the highest Silhouette Coefficient (0.059880). Although the corresponding Davies–Bouldin Index (1.452474) was not the lowest among all configurations, this combination of parameters provided the most meaningful clustering structure according to the Silhouette evaluation and was therefore selected as the final DBSCAN model.

Figures 10 and 11 show that the selected DBSCAN configuration produced 64 spatial clusters and identified 170 lightning observations as noise. Unlike the partition-based algorithms, DBSCAN generated clusters with irregular shapes and variable sizes according to the local density of lightning observations. Isolated observations that did not satisfy the density requirement were automatically classified as noise, enabling the algorithm to distinguish sparse observations from dense spatial groups.

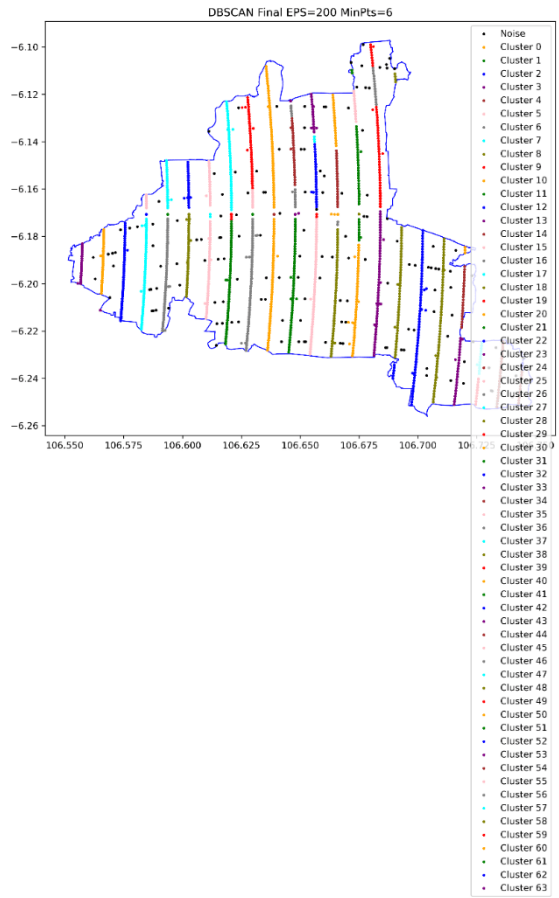


Figure 10. Spatial Distribution of CG+ lightning Clusters Generated using DBSCAN (ε = 200 m, MinPts = 6)

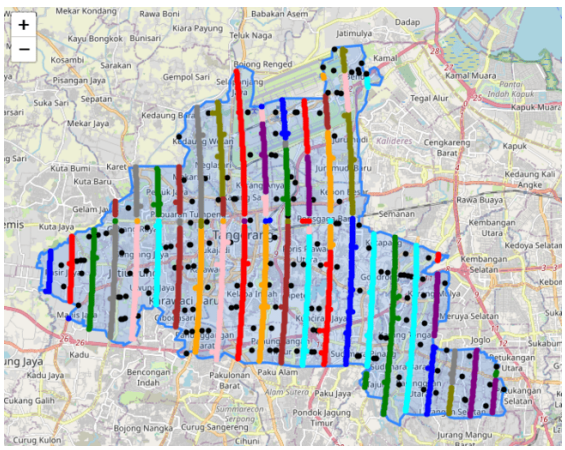


Figure 11. DBSCAN Clustering Results Overlaid on the OpenStreetMap basemap.

However, the resulting clusters are much more fragmented than those obtained using K-Means and K-

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Medoids. The large number of clusters indicates that the chosen density threshold separates many local groups of lightning observations, rather than forming a small number of broader spatial regions. As a result, although DBSCAN successfully detected density-based structures, the overall clustering quality was lower than that of partition-based algorithms according to internal validation metrics.

Performance Comparison of Clustering Algorithms

To objectively compare the clustering performance, the best results obtained from K-Means, K-Medoids, and DBSCAN were evaluated using the Silhouette Coefficient and the Davies–Bouldin Index (DBI). The best parameter configuration for each algorithm was selected based on the evaluation conducted in the previous sections. The comparison results are presented in Table 6.

Table 6. Performance Comparison of Clustering Algorithms

Algorithm	Best Parameter	Number of Clusters	Silhouette Coefficient	Davies–Bouldin Index
K-Means	K = 3	3	**0.445640**	**0.789620**
K-Medoids	K = 3	3	0.444728	0.794027
DBSCAN	ε = 200 m, MinPts = 6	64	0.059880	1.452474

.

Table 6 shows that the K-Means algorithm achieved the highest clustering performance among the three methods evaluated. This algorithm produced the highest Silhouette Coefficient (0.445640) and the lowest Davies–Bouldin Index (0.789620), indicating that the resulting clusters are more compact and better separated than those produced by the other algorithms.

The K-Medoids algorithm produced results very similar to those of the K-Means algorithm. With a Silhouette Coefficient of 0.444728 and a Davies–Bouldin Index of 0.794027, the differences between these two partition-based algorithms are relatively small. These findings suggest that the spatial distribution of CG+ lightning observations is sufficiently stable, allowing both centroid- and medoid-based clustering methods to produce consistent cluster structures.

In contrast, the DBSCAN algorithm produced a very different clustering structure. With the optimal parameter combination (ε = 200 m and MinPts = 6), DBSCAN generated 64 clusters and identified 170 observations as noise. Although this configuration achieved the highest Silhouette Coefficient among all DBSCAN experiments, its value (0.059880) remains significantly lower than that obtained using K-Means and K-Medoids. Similarly, the Davies–Bouldin Index (1.452474) is much higher, indicating lower cluster compactness and weaker separation.

These findings indicate that partition-based clustering is more suitable for describing the spatial distribution of CG+ lightning strikes in Tangerang City than density-based clustering. Rather than forming dense, isolated areas, the lightning observations were distributed fairly evenly throughout the study area. Consequently, while DBSCAN separated the data into many small local clusters based on point density, K-Means and K-Medoids were able to cluster the observations into three distinct spatial clusters.

The superior performance of K-Means can be explained by the spatial characteristics of CG+ lightning observations in Tangerang City. With few distinct high-density hotspots, the distribution of lightning strikes in the study area is generally continuous. Partition-based clustering performs better in situations like this because it divides the observations into dense, well-separated regional clusters based on overall geographic proximity. Therefore, K-Means can identify representative cluster centers that capture the dominant spatial structure of the dataset while maintaining high cluster density and separation.

On the other hand, DBSCAN uses local point density to identify clusters. The chosen density threshold often divides the dataset into many small clusters and a number of noise observations, since CG+ lightning observations are scattered across a relatively large area with varying local densities. Although this behavior allows DBSCAN to identify local spatial differences, it also fragments the overall distribution pattern, resulting in poorer separation and lower cluster cohesion compared to partition-based methods. Similarly, K-Medoids produces results similar to K-Means because there are few extreme outliers in the dataset, which reduces the benefit of selecting medoids rather than centroids as cluster representatives.

Overall, the experimental results show that K-Means is the most appropriate clustering model for describing the spatial distribution of Positive Cloud-to-Ground (CG+) lightning strikes in Tangerang City, while K-Medoids offers comparable performance with only a slight decrease in clustering quality. Although DBSCAN is effective at detecting local density variations and identifying noisy observations, its performance on this dataset is limited because the lightning observations exhibit a relatively continuous spatial distribution rather than clear density-based clustering.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Discussion

A comparative analysis shows that the choice of clustering algorithm has a significant impact on the spatial representation of CG+ lightning observations. Both K-Means and K-Medoids consistently identify three main spatial clusters, indicating that lightning observations can be effectively grouped into a small number of geographically coherent regions. The similarity in their evaluation scores suggests that the dataset contains relatively few extreme outliers, allowing both centroid-based and medoid-based approaches to produce nearly identical clustering structures.

In contrast, DBSCAN produces a much larger number of clusters because this algorithm clusters observations based on local point density rather than global partitioning. As a result, this algorithm successfully detects small local clusters and isolated observations; however, it also fragments the dataset into many clusters, thereby reducing the overall compactness and separation as measured by the Silhouette Coefficient and the Davies–Bouldin Index.

The superior performance of K-Means can be explained by the spatial characteristics of CG+ lightning observations in Tangerang City. Lightning events exhibit a relatively continuous spatial distribution across the entire study area, rather than forming numerous isolated hotspots with high density. Under these conditions, partition-based clustering is more effective because it groups observations into dense, well-separated regional clusters based on overall spatial proximity. As a result, K-Means is able to identify representative cluster centers (centroids) that reflect the dominant spatial distribution of the dataset. In contrast, DBSCAN places greater emphasis on local density relationships, which causes observations to be divided into many small clusters when local density varies across the study area. These findings suggest that the selection of a clustering algorithm should not only consider quantitative validation metrics but also the intrinsic spatial characteristics of the data.

From a practical standpoint, partition-based algorithms provide a simpler and more easily interpretable representation of the spatial distribution of CG+ lightning in Tangerang City. The three main clusters identified by K-Means and K-Medoids can facilitate the interpretation of regional lightning patterns and support land-use planning, infrastructure protection, and disaster mitigation strategies. Meanwhile, DBSCAN may be more suitable for applications requiring the identification of local density variations or isolated lightning events, rather than broad regional clustering.

The resulting geographic clusters provide useful information for infrastructure planning and disaster mitigation, although this study used internal validation measures to assess the performance of the clustering. Future research that incorporates external information such as population density, land use, critical infrastructure, or power transmission networks could utilize the identified clustering patterns as an initial geographic reference. By linking the spatial distribution of CG+ lightning with exposure and vulnerability characteristics, this integration will enable a more comprehensive evaluation of lightning risk. Consequently, the clustering results from this study serve as a foundation for future multi-source spatial analysis and decision-support applications.

The overall findings indicate that K-Means is the most suitable algorithm for clustering the spatial distribution of CG+ lightning observations in this study, as it achieves the best balance between cluster compactness and separation while maintaining a simple and easily interpretable cluster structure.

CONCLUSION

This study compares the performance of the K-Means, K-Medoids, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms in analyzing the spatial distribution of Positive Cloud-to-Ground (CG+) lightning observations in Tangerang City using lightning data collected throughout 2025. Prior to the clustering process, the dataset underwent data preprocessing, including data cleaning, CG+ observation selection, administrative boundary filtering, feature selection, coordinate transformation to the UTM coordinate system, and Min–Max normalization. This preprocessing stage resulted in a final dataset consisting of 10,235 valid CG+ lightning observations suitable for spatial clustering analysis.

The experimental results show that K-Means achieved the best clustering performance among the evaluated algorithms. Using K = 3, K-Means achieved the highest Silhouette Coefficient (0.445640) and the lowest Davies–Bouldin Index (0.789620), indicating the most compact and well-separated clusters. K-Medoids produced comparable results with a Silhouette Coefficient of 0.444728 and a Davies–Bouldin Index of 0.794027, indicating that both partition-based algorithms are effective in representing the spatial distribution of CG+ lightning observations. In contrast, DBSCAN produced a greater number of density-based clusters and successfully identified isolated observations as noise. However, its clustering performance was lower than that of the partition-based

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

algorithms based on the evaluation metrics used in this study. Overall, these findings indicate that K-Means is the most suitable algorithm for clustering the spatial distribution of CG+ lightning observations in Tangerang City. To further assess the beneficial implications of these clustering results for urban disaster risk management and infrastructure resilience, future research is recommended to integrate the identified lightning clusters with external spatial data, such as land use, population distribution, and critical infrastructure.

DAFTAR PUSTAKA

Agresto, J., Duan, Z., Bentley, M., Gerken, T., & Bonsal, D. (2024). Spatiotemporal-DBSCAN: A Density-Based Clustering Method for Analyzing Spatiotemporal Ground-lightning Dataset. *Conference Proceedings of the IEEE International Performance, Computing, and Communications Conference*. https://doi.org/10.1109/IPCCC59868.2024.10850463

Ashari, I. F., Dwi Nugroho, E., Baraku, R., Novri Yanda, I., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. *Journal of Applied Informatics and Computing*, *7*(1), 89–97. https://doi.org/10.30871/jaic.v7i1.4947

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. www.aaai.org

Han, J., Kamber, M., Pei, J., & Kaufmann, M. (2012). *Data Mining: Concepts and Techniques Third Edition*.

Jiang, R., Zhang, G., Wang, S., Xue, B., Xie, Z., Yu, T., Wang, K., Ding, J., & Zhu, X. (2023). Characteristics of cloud-to-ground lightning (CG) and differences between +CG and −CG strokes in China regarding the China National Lightning Detection Network. *Natural Hazards and Earth System Sciences*, *23*(12), 3747–3759. https://doi.org/10.5194/nhess-23-3747-2023

Karney, C. F. F. (2012). *Algorithms for geodesics*. https://doi.org/10.1007/s00190-012-0578-z

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data*. Wiley. https://doi.org/10.1002/9780470316801

Nugrahanto, Z. Z., & Yunin Sofro, A. (2025). Implementation of DBSCAN and K-MEANS++ Methods for Flood Vulnerability Cluster Mapping in East Java Province, 2024://creativecommons.org/licenses/by-sa/4.0). *Matematika Murni Dan Aplikasi*, *10*(2), 427–440. https://doi.org/10.18860/cauchy.v10i2.32446

Paski, J. A. I., Permana, Y. H., & Pertiwi, yah A. S. (2017). Analisis Sebaran Petir Cloud To Ground (CG) di Wilayah Jabodetabek Pada Tahun 2016. *Prosiding Seminar Nasional Fisika (E-Journal) SNF2017 UNJ*. https://doi.org/10.21009/03.SNF2017.02.EPA.10

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salman, F., & Fauziah, F. (2023). Comparison Analysis of K-Means and DBSCAN Algorithms for Improving Budget Absorption Efficiency in EIS. *Brilliance: Research of Artificial Intelligence*, *3*(2), 378–383. https://doi.org/10.47709/brilliance.v3i2.3373

Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems*, *101*. https://doi.org/10.1016/j.is.2021.101804

Shi, M., Fan, P., Tao, H., Li, Q., Wang, J., Liu, Y., & Wei, L. (2025). A lightning cluster identification method considering multi-scale spatiotemporal neighborhood relationships. *PLOS ONE*, *20*(10 October). https://doi.org/10.1371/journal.pone.0333207

Wang, J., Huang, Q., Ma, Q., Chang, S., He, J., Wang, H., Zhou, X., Xiao, F., & Gao, C. (2020). Classification of VLF/LF lightning signals using sensors and deep learning methods. *Sensors (Switzerland)*, *20*(4). https://doi.org/10.3390/s20041030

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.