CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 32-44
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Implementation of DBSCAN and K-MEANS++ Methods for Flood Vulnerability Cluster Mapping in East Java Province, 2024

Zalfa Zaliana Nugrahanto[1] and A'yunin Sofro[2*]

[1] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya*
[2] *Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya*

Abstract

Flood vulnerability in East Java varies across districts due to differences in hydrometeorological pressure and exposure levels. This study compares two clustering algorithms—DBSCAN and K-Means++—for identifying patterns in eleven flood-impact indicators. DBSCAN parameter selection was conducted using a k-distance graph, resulting in $\varepsilon = 0.8$ and $MinPts = 3$, which produced five clusters and three noise points. The Silhouette Index for DBSCAN was 0.3266, calculated including noise points to ensure fair evaluation against K-Means++, which obtained a Silhouette Index of 0.2453 for five clusters. The findings indicate that DBSCAN produced higher internal cohesion under the given dataset. However, the resulting clusters are not interpreted as validated flood risk zones or as physically causal patterns, due to the absence of external validation layers such as historical flood maps, hydrological data, or topographic information. The results therefore provide a methodological comparison between density-based and centroid-based clustering for flood-impact variables without making geographical or causal inferences.

Keywords: Flood vulnerability, DBSCAN, K-Means++, Clustering, East Java Province, Silhouette Index.

1 Introduction

Flood hazard research increasingly incorporates machine learning and clustering techniques to analyze multidimensional disaster indicators [1], [2]. Although numerous studies have demonstrated the practical usefulness of unsupervised learning—for example, density-based clustering for early warning systems [3] the majority of prior works prioritize empirical mapping rather than examining whether the selected algorithms are structurally compatible with the statistical properties of hydrometeorological data. Consequently, the literature is dominated by case-specific implementations with limited methodological evaluation or cross-algorithm comparison.

Partition-based methods such as K-Means and K-Means++ remain popular in flood susceptibility analysis due to their computational efficiency [4], [5]. However, existing studies seldom acknowledge that these algorithms assume convex cluster structures and rely on Euclidean distance as their similarity metric, which leads to instability in the presence of outliers, skewed distributions, or extreme-magnitude events—characteristics frequently observed in disaster-impact

[*] **Corresponding author**. E-mail: ayuninsofro@unesa.ac.id

Submitted: November 9, 2025 Reviewed: November 22, 2025 Accepted: December 12, 2025
DOI: https://doi.org/10.18860/cauchy.v11i1.37410

datasets [6], [7]. These limitations are rarely scrutinized, resulting in analyses whose clustering structure may be strongly influenced by noise rather than underlying hazard patterns.

Density-based methods such as DBSCAN offer a contrasting clustering paradigm by allowing arbitrarily shaped clusters and explicitly identifying noise [8], [9]. Several studies highlight DBSCAN's theoretical suitability for environmental and hazard data containing heterogeneous densities or localized anomalies [10]. However, much of this evidence remains conceptual: prior research does not rigorously evaluate how DBSCAN behaves relative to centroid-based methods under controlled parameter selection, nor does it examine how noise treatment affects internal validity metrics. Comparative DBSCAN–K-Means++ studies remain scarce, and those that exist are predominantly situated in industrial or digital analytics domains [11], [12], limiting their relevance for flood-related applications.

A further limitation in the existing flood-clustering literature is the reliance on single-algorithm analyses [5], [8]. Such studies typically report cluster results without discussing algorithmic assumptions, sensitivity to parameter choices, or fairness in comparing methods with differing noise-handling mechanisms. This absence of methodological scrutiny restricts the interpretability of results and obscures whether observed cluster differences arise from genuine hazard patterns or from algorithm-specific biases.

In this context, East Java serves not as a source of novelty, but as a suitable testbed because its district-level flood-impact indicators exhibit high variance, localized extremes, and irregular density distributions [13], [14]. These characteristics allow for a meaningful assessment of how structurally different clustering paradigms respond to disaster-impact data.

Accordingly, this study fills the methodological gap by conducting a controlled comparison between DBSCAN and K-Means++ using a unified internal validity metric the Silhouette Index—while applying consistent preprocessing, fair parameter selection, and transparent noise treatment. The aim is not to infer geographical causation or identify physical patterns in flood-impact indicators, but to clarify how density-based and centroid-based algorithms behave when exposed to heterogeneous hazard-impact data. This methodological clarification is essential for improving the reliability, reproducibility, and interpretability of clustering-based flood assessments.

2 Methods

This section outlines the methodological framework employed to compare DBSCAN and K-Means++ for clustering flood-impact indicators in East Java. The research procedure encompasses five sequential stages: data acquisition and preparation, standardization, application of the DBSCAN algorithm, application of the K-Means++ algorithm, and cluster validation. Each stage is designed to ensure a fair and reproducible comparison between the two clustering paradigms while maintaining consistency in preprocessing and evaluation.

2.1 Research Procedure

The research was conducted systematically through several stages, beginning with a detailed description and preparation of the dataset used in the analysis.

The dataset consists of flood-impact records from 38 districts/cities in East Java Province, compiled from the Indonesian Meteorological Agency (BMKG), the Geospatial Information Agency (BIG), and the Central Bureau of Statistics (BPS). All data correspond to the 2024 reporting period and use district/city (ADM2) administrative polygons as the spatial analysis unit. For each spatial unit, a set of eleven disaster-impact indicators (X1–X11) was assembled to represent the multidimensional consequences of flood events.

These eleven disaster impact indicators consist of the number of flood events (X1), which represents the annual frequency of flood occurrences in each district or city. Fatalities and missing persons (X2) indicate the total number of individuals reported dead or missing due to flood events. Injured or sick victims (X3) count individuals who suffered injuries or illnesses as a direct result of flooding. Evacuated and affected population (X4) measures the number of residents displaced or otherwise affected by flooding.

In addition, housing damage is represented by severely damaged houses (X5), which quantify residential buildings categorized as severely damaged. Moderately damaged houses (X6) enumerate residential buildings classified as moderately damaged. Slightly damaged houses (X7) document houses that sustained minor structural damage. Inundated houses (X8) count residential units flooded without structural collapse.

Infrastructure damage is captured through damaged educational facilities (X9), which record schools or educational facilities impacted by flood events. Damaged religious facilities (X10) note religious buildings affected by flooding. Damaged health facilities (X11) enumerate hospitals, clinics, or health centers impacted by flood events.

All indicators were entered into the analysis system and standardized using the *z-score* transformation to ensure uniform measurement scales and to minimize bias among variables.

Multicollinearity diagnostics were not included in the final analysis pipeline, as this study focuses on unsupervised clustering of standardized flood-impact indicators rather than parameter estimation or inferential modeling.

The DBSCAN algorithm was applied by calculating pairwise distances using the *Euclidean Distance*. The parameters ε and *MinPts* were determined using the k-distance graph derived from the *K-Nearest Neighbors (KNN)* plot to identify density-based clusters and noise points.

The number of clusters (K) for the centroid-based approach was determined based on exploratory examination of spatial patterns and supported by analytical cluster tendency. The K-Means++ algorithm was then used to initialize centroids efficiently, followed by iterative recalculation until convergence, producing stable and distinct cluster groupings.

Clustering validity from both DBSCAN and K-Means++ was assessed using the *Silhouette Index* (SI), reflecting cohesion and separation of cluster structures. The resulting clusters were examined using spatial visualization to illustrate the distribution of indicator based patterns, without implying geographical causation or physical flood mechanisms.

2.2 Data Standardization

Before performing the clustering analysis, all research variables were standardized to eliminate the effect of different measurement scales [15]. Standardization was carried out using the *Z-score* transformation, which converts each variable into a common scale with a mean of zero and a standard deviation of one [16].

The *Z-score* for each observation Z_{ij} is calculated as follows:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j} \tag{1}$$

where X_{ij} represents the value of variable j for observation i, $\bar{X}_j$ is the mean of variable j, and s_j is the standard deviation of variable j.

This transformation ensures that variables with larger numerical ranges do not dominate the clustering process, allowing all attributes to contribute equally to the distance calculations. The standardized data were subsequently used as input for both the DBSCAN and K-Means++ clustering methods.

2.3 DBSCAN

The clustering process was performed using the *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) algorithm, which groups data points based on density similarity rather than predefined cluster numbers [17]. DBSCAN is particularly effective for identifying arbitrary-shaped clusters and distinguishing noise or outlier data [18].

DBSCAN requires two key parameters: the neighborhood radius (ε) and the minimum number of points (*MinPts*) required to form a dense region [10]. The distance between two data points was measured using the *Euclidean Distance*, formulated as:

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

where x_i and x_j denote two observations in a p-dimensional space. A point x_i is considered a *core point* if the number of data points within its ε-neighborhood is at least equal to *MinPts*, defined as:

$$N_\varepsilon(x_i) = \{x_j \in D \mid d(x_i, x_j) \leq \varepsilon\} \tag{3}$$

The algorithm iteratively expands clusters by connecting density-reachable points until no further expansion is possible. Points that do not belong to any cluster are classified as noise. Parameter selection for ε and *MinPts* was performed using the k-distance graph generated from the *K-Nearest Neighbors (KNN)* distance plot, which provides a clearer identification of density thresholds. The resulting clusters were then visualized spatially using QGIS to observe the distribution patterns of flood-impact indicators.

2.4 K-Means++

The K-Means++ algorithm was employed to perform partition-based clustering by minimizing the total variance within clusters [19]. This method improves upon the conventional K-Means by optimizing the initialization of cluster centroids, thereby enhancing convergence stability and reducing sensitivity to initial random selection.

The algorithm begins by selecting the first centroid randomly from the dataset. Subsequent centroids are chosen probabilistically, where data points farther from existing centroids have a higher probability of being selected [20]. The probability for selecting a data point x_i as the next centroid is defined as:

$$P(x_i) = \frac{D(x_i)^2}{\sum_{k=1}^{n} D(x_k)^2} \tag{4}$$

where $D(x_i)$ denotes the shortest Euclidean distance between x_i and the nearest existing centroid.

Once K centroids are initialized, each data point is assigned to the nearest centroid using the *Euclidean Distance*:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - c_{jk})^2} \tag{5}$$

After assignment, new centroids are recalculated as the mean of all points belonging to each cluster:

$$c_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i \tag{6}$$

where n_j represents the number of points in cluster C_j. The algorithm iteratively updates cluster memberships and centroid positions until convergence is achieved, indicated by minimal or no change in centroid values between iterations.

This process yields compact and well-separated clusters. The resulting patterns were then compared with DBSCAN outcomes to evaluate the effectiveness of each method in identifying spatial distributions of patterns in flood-impact indicators.

2.5 Cluster Validation and Evaluation

The quality of the cluster structures generated by DBSCAN and K-Means++ was assessed using the Silhouette Index (SI). This metric provides a unified measure of clustering validity by quantifying both the internal cohesion of data points within a cluster and their separation from neighboring clusters. Higher SI values reflect well-defined and distinct clusters, whereas values approaching zero or negative indicate weak cohesion or overlap between cluster boundaries.

2.5.1 Silhouette Index (SI)

The Silhouette Index evaluates both cohesion (how close data points are within a cluster) and separation (how far they are from other clusters). For each observation i, the Silhouette coefficient $s(i)$ is computed as [21]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{7}$$

where $a(i)$ is the average distance between observation i and all other points within the same cluster, $b(i)$ is the minimum average distance between observation i and all points belonging to other clusters. The overall Silhouette Index is the average of all $s(i)$ values:

$$SI = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{8}$$

The SI value ranges between -1 and 1, where values close to 1 indicate well-formed clusters, values near 0 suggest overlapping clusters, and negative values imply misclassification.

3 Results and Discussion

The results are structured into five components: data preprocessing outcomes, DBSCAN clustering results, K-Means++ clustering results, comparative algorithm evaluation, and spatial pattern interpretation. Each component is presented with a focus on methodological insights rather than geographical hazard claims, aligning with the study's comparative objective.

3.1 Evaluation of DBSCAN Parameters and Cluster Formation

To identify the optimal values of ε and *MinPts*, a k-distance graph was constructed using the *K-Nearest Neighbors (KNN)* distance plot. The k-distance curve increases gradually at first, and the optimal value of ε appears at the point where the slope begins to rise sharply. This inflection point indicates the transition from dense to sparse regions in the dataset. The resulting k-distance graph is presented in Figure 1.

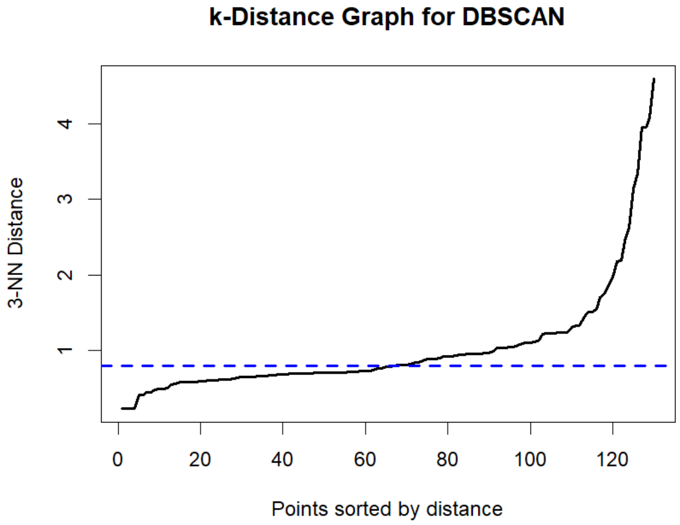

Figure 1: k-distance graph used to determine the optimal value of ε for DBSCAN.

Based on the k-distance graph shown in Figure 1, the curve displays a clear inflection around $\varepsilon = 0.8$, which was selected as the optimal neighborhood radius, with *MinPts* fixed at 3. To achieve this, several combinations of ε and *MinPts* were tested using the *Silhouette Index* as the primary validity measure. A higher Silhouette value indicates that data objects are well grouped within their clusters, while a value close to zero or negative indicates less compact cluster structures or possible overlap between clusters.

The evaluation results of the parameter combinations are presented in Table 1. Based on the evaluation, the highest *Silhouette Score* was obtained at $\varepsilon = 0.8$ and $MinPts = 3$, with a score of 0.3266. This result shows that the selected parameters provide the best clustering configuration among all combinations tested.

Table 1: Evaluation Results of DBSCAN Parameter Combination using Silhouette Index

Epsilon (ε)	**MinPts**	**Silhouette Score**
0.2	3	0.0134
0.2	4	-0.1696
0.2	5	-0.1696
0.2	6	-0.1696
0.3	3	0.0943
0.3	4	-0.0866
0.4	3	0.1818
0.4	4	0.1818
0.5	3	0.1818
0.6	3	0.2403
0.6	4	0.2403
0.6	5	0.2403
0.7	3	0.2403
0.7	4	0.2403
0.8	3	**0.3266**
0.8	4	0.2729
0.9	3	0.3266
0.9	4	0.3008
1.0	3	0.3266
1.0	4	0.3266

Based on the parameter optimization results, the DBSCAN algorithm was then implemented using $\varepsilon = 0.8$ and $MinPts = 3$, which produced a total of five (5) distinct clusters and three (3) noise points. The formation of these clusters indicates differences in flood intensity, affected populations, and infrastructure damage among the districts and cities analyzed.

The results of the clustering process are summarized in Table 2, which shows the composition of each cluster and its respective regions. In general, Cluster 0 represents districts characterized by higher values across multiple flood impact indicators, particularly those related to affected population and infrastructure damage, Cluster 1 represents districts exhibiting moderate to low values across flood impact indicators while Clusters 2–4 exhibit distinct combinations of flood-impact indicators (X1–X11), characterized by differing magnitudes of affected population, housing damage, and infrastructure disruption, without inferring specific geographical or hydrological causes.

The results demonstrate that DBSCAN successfully identified coherent groupings of patterns in flood-impact indicators regions across East Java Province. The spatial visualization of these clusters is intended solely to illustrate indicator based patterns and to facilitate methodological comparison, rather than to serve as a basis for operational flood risk management or policy related applications.

Based on the results of the DBSCAN clustering analysis, the spatial distribution of patterns in flood-impact indicators in East Java Province can be visualized as shown in Figure 2. This map illustrates the division of districts and cities into several clusters according to their flood

Table 2: List of Districts/Cities Based on DBSCAN Clustering Results

Cluster	Districts/Cities
Cluster 0	Kabupaten Ponorogo, Kabupaten Blitar, Kabupaten Lumajang, Kabupaten Jember, Kabupaten Situbondo, Kabupaten Pasuruan, Kabupaten Sidoarjo, Kabupaten Mojokerto, Kabupaten Jombang, Kabupaten Madiun, Kota Surabaya
Cluster 1	Kabupaten Trenggalek, Kabupaten Kediri, Kabupaten Banyuwangi, Kabupaten Probolinggo, Kabupaten Magetan, Kabupaten Ngawi, Kabupaten Nganjuk, Kota Madiun, Kota Mojokerto, Kota Blitar, Kota Kediri, Kabupaten Pacitan, Kabupaten Tulungagung, Kabupaten Bangkalan, Kabupaten Pamekasan, Kabupaten Sumenep, Kota Batu Kota Probolinggo, Kabupaten Malang
Cluster 2	Kabupaten Bondowoso
Cluster 3	Kabupaten Bojonegoro, Kota Pasuruan
Cluster 4	Kabupaten Gresik, Kabupaten Lamongan

characteristics, which were formed based on variables such as the number of flood incidents, affected population, and infrastructure damage.

Each color in the map corresponds to a cluster group, namely: Cluster 0 (green), Cluster 1 (orange), Cluster 2 (blue), Cluster 3 (pink), and Cluster 4 (light green). Districts and cities that are not assigned to any cluster (noise) are excluded from the visualization for clarity.

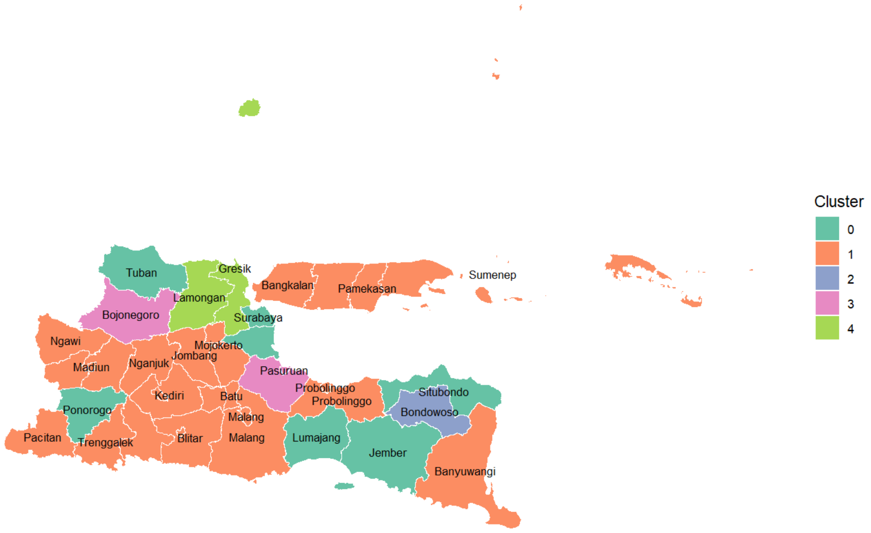

Figure 2: Map of cluster groupings generated by the DBSCAN algorithm

As illustrated in Figure 2, the cluster distribution illustrates variations in indicator based groupings across administrative units in East Java. Regions classified into **Cluster 0** (green) exhibit the highest magnitudes of flood-related impact indicators, including Jember, Lumajang, Pasuruan, and Sidoarjo. **Cluster 1** (orange) encompasses districts with moderate values of flood related impacts and limited infrastructure disruption, such as Kediri, Trenggalek, and Ngawi. **Cluster 2** and **Cluster 3** representdistricts with moderate flood-impact indicator values, characterized by intermediate levels of affected population and housing damage, including Bondowoso, Bojonegoro, and Kota Pasuruan. Meanwhile, Cluster 4 (light green) consists of districts characterized by relatively higher numerical values in specific flood-impact indicators, particularly the number of inundated houses and housing damage. These patterns reflect similarities in the numerical values of flood impact indicators across districts, rather than similarities in geographical settings or flood mechanisms.

The spatial distribution of the clusters illustrates differences in flood-impact indicator profiles across administrative units in East Java. The observed cluster patterns are derived exclusively from similarities in the reported flood-impact indicators and should not be interpreted as evidence of underlying geographical settings or physical flood mechanisms. Instead, the results highlight variations in the magnitude and combination of impacts—such as affected population and infrastructure damage—as captured in the dataset. Accordingly, the cluster based spatial analysis is intended to identify indicator based patterns in flood impacts and to provide preliminary insights that may inform further investigation when complemented by additional data and external validation, rather than to delineate validated flood hazard or risk zones.

After identifying the DBSCAN structure and its optimal parameters, the next stage applies the K-Means++ algorithm to the same standardized dataset to evaluate whether a centroid-based method produces comparable patterns

3.2 Clustering Analysis Using K-Means++ Algorithm

The K-Means++ algorithm was utilized to perform clustering on the flood disaster data of East Java Province. This algorithm is an improved version of the traditional K-Means method, providing better initialization of cluster centroids and minimizing the possibility of suboptimal clustering results (Arthur & Vassilvitskii, 2007). Before clustering, the dataset was standardized using the Z-score normalization method to ensure that all variables contributed equally, as each variable had different measurement scales and magnitudes.

3.2.1 Determination of the Optimal Number of Clusters (Elbow Method)

To determine the optimal number of clusters (k), the *Elbow Method* was applied. This method examines the relationship between the number of clusters and the total within-cluster sum of squares (WSS). The WSS value decreases as k increases, but after a certain point, the rate of decrease slows down, forming an "elbow" shape. This point represents the optimal number of clusters, balancing accuracy and model simplicity.

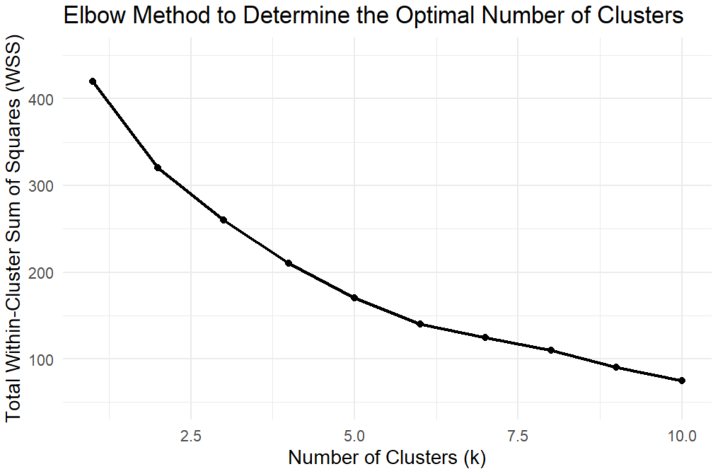

Figure 3: Elbow

As illustrated in Figure 3, the inflection point occurs at $k = 5$, indicating that five clusters provide the best trade-off between compactness and separation. Hence, the number of clusters for this study was set to five.

To determine the appropriate number of clusters for the K-Means++ algorithm, the Elbow Method was initially used, as illustrated in Figure 3. While the Within-Cluster Sum of Squares (WSS) curve shows a gradual decline, an observable inflection point appears around $k = 5$. However, because the elbow was not sharply defined, additional validation was conducted using the Silhouette Index (SI) to strengthen the selection of the optimal number of clusters.

After selecting $k = 5$, the K-Means++ algorithm was executed using the following parameters: `num_init = 10`, `max_iters = 100`, and `initializer = "kmeans++"`. Each observation was assigned to the cluster whose centroid had the minimum Euclidean distance. The algorithm iteratively minimized the total within-cluster variance until convergence was achieved.

3.2.2 Cluster Formation Using K-Means++

After determining the optimal number of clusters, the K-Means++ algorithm was applied to the standardized dataset using the parameters num_init = 10 and max_iters = 100. The clustering results indicate that Cluster 1 consists of 28 districts/cities with lower to moderate values across most flood impact indicators, including relatively small numbers of flood events, affected populations, and housing damage. Cluster 2 comprises 5 districts/cities showing moderate levels of flood-related impacts, characterized by increases in affected population (X4), housing damage (X5–X7), and inundated houses (X8).

Cluster 3 includes 3 districts/cities that exhibit higher values in certain indicators, particularly in terms of affected populations (X4) and inundated houses (X8), reflecting more significant impacts on residents and housing. In contrast, Cluster 4 consists of a single district/city that shows higher values in structural damage indicators (X5–X7) and public facility disruptions (X9–X11), indicating a distinct pattern of infrastructure-related impacts. Finally, Cluster 5 represents one district/city experiencing extreme values in housing destruction (X5–X7) and inundated houses (X8), highlighting a localized area with very high impact levels despite its more concentrated affected area.

The resulting cluster label was then appended to the original dataset and exported as an Excel file named *Hasil_Cluster_Banjir_JawaTimur.xlsx*. Table 3 summarizes the average values of each numerical variable within each cluster.

Table 3: Average Values of Each Variable by Cluster (K-Means++, 5 Clusters)

Cluster	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	2.412	0.092	0.014	4870.33	0.128	0.000	1.667	812.45	0.041	0.082	0.000
2	3.876	0.221	0.087	10542.12	2.143	1.879	4.221	1924.67	0.067	0.054	0.032
3	10.333	0.667	1.000	48160.67	0.000	0.000	0.000	7603.00	0.333	0.000	0.333
4	6.459	0.503	0.421	21673.45	4.887	3.214	7.558	3409.33	0.188	0.091	0.061
5	2.000	2.000	0.000	6405.00	10.000	10.000	16.000	1431.00	0.000	0.000	0.000

3.2.3 Outlier Sensitivity in K-Means++ Results

Although the K-Means++ algorithm successfully partitioned the dataset into five clusters, the results reveal a notable sensitivity to outliers. Because K-Means++ does not include a mechanism for separating extreme observations from the main data structure, all districts—including those with unusually high impact values—were forced into one of the predefined clusters.

This behavior resulted in two *singleton clusters* (Cluster 4 and Cluster 5 in Table 3) in the final clustering result, each consisting of a single district that strongly influenced centroid computation due to extreme values. The presence of these outliers increased within-cluster variance and directly contributed to the lower Silhouette Index (0.2453) observed for K-Means++.

These findings indicate that the reduced validity of K-Means++ is not only a result of its Euclidean and convex-cluster assumptions, but also its inability to isolate irregular or extreme-impact districts—an issue that DBSCAN handles more effectively by assigning such observations as noise.

3.3 Comparison Based on Silhouette Index

The Silhouette Index is a measure of how well each object lies within its cluster, with values ranging from -1 to 1. A higher Silhouette Index indicates that the data points are well matched to their own cluster and poorly matched to neighboring clusters, signifying better-defined and more cohesive groups.

Table 4: Comparison of Silhouette Index Values Between DBSCAN and K-Means++

Method	**Silhouette Index (↑)**
DBSCAN	0.3266
K-Means++	0.2453

As shown in Table 4, the *DBSCAN* algorithm achieved a higher Silhouette Index value (0.3266) compared to *K-Means++* (0.2453). This result indicates that DBSCAN forms clusters that are more cohesive and better separated than those produced by K-Means++. Therefore, it can be concluded that, based on the Silhouette Index criterion, **DBSCAN provides superior clustering performance** for identifying patterns in flood-impact indicators in East Java Province.

3.4 Average Characteristics of Each Cluster (DBSCAN Results)

Table 5 presents the average values of all numerical flood indicators for each cluster generated by the *DBSCAN* algorithm. These averages represent the general characteristics of each group of districts or cities in East Java, classified according to similarities in their flood impact indicator values

Table 5: Average Values of Each Variable by Cluster (DBSCAN)

Cluster	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**X10**	**X11**
0	5.182	0.545	0.364	18617.82	1.455	0.909	7.636	2760.18	0.273	0.182	0.091
1	1.476	0.000	0.000	3259.14	0.000	0.000	0.000	704.05	0.000	0.000	0.000
2	2.500	0.000	0.000	690.00	0.000	0.000	0.000	186.50	0.000	1.000	0.000
3	6.000	1.000	0.000	20494.50	0.000	0.000	0.000	1341.50	0.000	0.000	0.000
4	8.000	0.000	0.000	4273.00	0.000	0.000	0.000	3241.50	0.000	0.000	0.000

Based on Table 5, the DBSCAN algorithm produced five distinct clusters representing different levels of flood impact across East Java. **Cluster 0** shows the highest flood impact, with dominant values in X4 (affected population) and X8 (inundated houses), representing regions consistently exposed to severe and widespread flooding. **Cluster 1** contains districts with low averages across all indicators, indicating areas experiencing light or infrequent flooding. **Cluster 2** represents localized flood events, characterized by generally low values with a minor anomaly in X10 (damaged religious facilities). **Cluster 3** exhibits moderately high values, indicating areas with significant population impact (X4) and infrastructure disruption. **Cluster 4** contains districts with substantial damage to residential structures, reflected by high values in X1 (flood events) and X8 (total inundation).

4 Conclusion

This study conducted a methodological comparison between DBSCAN and K-Means++ in clustering district level flood impact indicators in East Java. Using a unified preprocessing pipeline and a fairness controlled evaluation, DBSCAN achieved a higher Silhouette Index (0.3266) than K-Means++ (0.2453) when noise points were included in the computation. This indicates that the density-based approach produces more coherent and internally consistent clusters when the data exhibit irregular distributions and localized extremes.

The findings highlight that DBSCAN is better suited for multivariate disaster impact datasets characterized by heterogeneous magnitudes and non-linear structures, whereas K-Means++ shows reduced stability under the same conditions due to its reliance on Euclidean distance and convex-cluster assumptions. The results therefore clarify algorithmic behavior rather than geographical susceptibility, and the interpretation is restricted to methodological performance rather than spatial hazard inference.

Several limitations remain in this study. First, the analysis relies on a single-year dataset, which may not capture interannual variability in flood impacts. Second, only flood-impact indicators were used, excluding environmental and infrastructural variables that could provide more comprehensive context. Third, and importantly, the spatial interpretation of clusters remains qualitative without quantitative validation through overlay analysis with external geographical data such as topographic maps, river networks, or independent flood hazard assessments. Consequently, the observed spatial patterns should be interpreted as indicator-based groupings rather than validated flood risk zones.

Future work may address these limitations by incorporating multi-year temporal patterns, integrating external validation layers, exploring hybrid clustering approaches, and expanding variable sets to improve analytical robustness. Overall, this study contributes a transparent and reproducible methodological comparison that can support more informed algorithm selection in clustering-based flood impact analysis.

CRediT Authorship Contribution Statement

Zalfa Zaliana Nugrahanto: Conceptualization, Methodology, Software, Formal Analysis, Data Curation, Writing – Original Draft, Visualization. **A'yunin Sofro:** Conceptualization, Methodology, Validation, Writing – Review & Editing, Supervision, Project Administration.

Declaration of Generative AI and AI-assisted technologies

No generative AI or AI-assisted technologies were used during the preparation of this manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The authors gratefully acknowledge the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG), the Geospatial Information Agency (BIG), and the Central Bureau of Statistics (BPS) for providing the official flood-impact data used in this study. All data correspond to the 2024 reporting period and use district/city (ADM2) administrative polygons as the spatial analysis unit, as provided by BIG.

Data and Code Availability

The dataset and code supporting the findings of this study are available from the corresponding author upon reasonable request. Flood impact data were obtained from official Indonesian government agencies and are subject to their data sharing policies.

References

[1] Z. Yuan et al., “Dynamic evolution and scenario-based prediction of urban flood resilience: A system dynamics modeling approach in kunming, china,” *Journal of Environmental Management*, vol. 395, p. 127 740, 2025. DOI: 10.1016/j.jenvman.2025.127740. Available online.

[2] M. F. Ghifari, K. Rusba, and M. Ramdan, “Kebijakan penanggulangan bencana banjir dan kebakaran di kota balikpapan,” *Identifikasi*, vol. 10, no. 1, pp. 156–160, 2024.

[3] J. L. Duykers, K. Ardana, R. Yulistiani, E. Irwansyah, and D. Fitrianah, “Identifying factors for supporting early warning flood using clustering approach and geo-spatial analysis,” *Procedia Computer Science*, vol. 227, pp. 540–547, 2023. DOI: 10.1016/j.procs.2023.10.556.

[4] E. A. Okoli et al., “Integrated flood susceptibility mapping using machine learning and geospatial techniques: A case study of imo state, southeastern nigeria,” *Journal of African Earth Sciences*, vol. 233, p. 105 872, 2026. DOI: 10.1016/j.jafrearsci.2025.105872.

[5] H. Xu, C. Ma, J. Lian, K. Xu, and E. Chaima, “Urban flooding risk assessment based on an integrated K-Means cluster algorithm and improved entropy weight method in the region of haikou, china,” *Journal of Hydrology*, vol. 563, pp. 975–986, 2018. DOI: 10.1016/j.jhydrol.2018.06.060.

[6] A. A. Abdulnassar and L. R. Nair, “Performance analysis of K-Means with modified initial centroid selection algorithms and developed K-Means9+ model,” *Measurement: Sensors*, vol. 25, p. 100 666, 2023. DOI: 10.1016/j.measen.2023.100666.

[7] S. K. Majhi and S. Biswal, “Optimal cluster analysis using hybrid K-Means and ant lion optimizer,” *Karbala International Journal of Modern Science*, vol. 4, no. 4, pp. 347–360, 2018. DOI: 10.1016/j.kijoms.2018.09.001.

[8] Y. Tu, Z. Tang, and B. Lev, “Regional flood risk grading assessment considering indicator interactions among hazard, exposure, and vulnerability: A novel FlowSort with DBSCAN,” *Journal of Hydrology*, vol. 639, p. 131 587, 2024. DOI: 10.1016/j.jhydrol.2024.131587.

[9] A. M. Jibrin, M. Al-Suwaiyan, Z. M. Yaseen, and S. I. Abba, “New perspective on density-based spatial clustering of applications with noise for groundwater assessment,” *Journal of Hydrology*, vol. 661, p. 133 566, 2025. DOI: 10.1016/j.jhydrol.2025.133566.

[10] O. Kulkarni and A. Burhanpurwala, “A survey of advancements in dbscan clustering algorithms for big data,” in *2024 3rd International conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC)*, IEEE, 2024, pp. 106–111. DOI: 10.1109/PARC59193.2024.10486339.

[11] F. M. Martínez-García, A. Molina García, F. C. Gómez de León, and M. Alarcón, “Distillation anomaly and fault detection based on clustering algorithms,” *Journal of Industrial Information Integration*, vol. 48, p. 100 970, 2025. DOI: 10.1016/j.jii.2025.100970.

[12] S. Srinath and N. Baydeti, “Behavioral pattern clustering for thematic user segmentation in web interaction environments,” *Information Sciences*, vol. 724, p. 122 745, 2026. DOI: 10.1016/j.ins.2025.122745.

[13] B. E. Cahyono, E. Purwandari, Misto, N. Febrianti, and Y. Pramono, “Mapping flooded risk area in east java indonesia using remote sensing data,” in *Journal of Physics: Conference Series*, vol. 1825, Bristol, UK: IOP Publishing, 2021, p. 012 081. DOI: 10.1088/1742-6596/1825/1/012081. Available online.

[14] Y. S. Iswandaru, A. Widodo, and M. S. Purwanto, “Analysis of flood disaster areas, jombang regency, east java using arcgis remote sensing,” *Journal of Marine-Earth Science and Technology*, vol. 2, no. 3, pp. 78–82, 2021. DOI: 10.12962/j27745449.v2i3.100. Available online.

[15] P. P. Allorerung, A. Erna, M. Bagussahrir, and S. Alam, “Analisis performa normalisasi data untuk klasifikasi k-nearest neighbor pada dataset penyakit,” *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, pp. 178–191, 2024. DOI: 10.14421/jiska.2024.9.3.178-191.

[16] I.-K. Hong et al., “Investigating standardized criteria to evaluate the impact of agro-healing programs on psychological and interpersonal outcomes: Utilizing normalization methods,” *Acta Psychologica*, vol. 261, p. 105 727, 2025. DOI: 10.1016/j.actpsy.2025.105727.

[17] F. Gani, H. S. Panigoro, S. L. Mahmud, E. Rahmi, S. K. Nasib, and L. O. Nashar, “Implementation of k-nearest neighbor algorithm on density-based spatial clustering application with noise method on stunting clustering,” *Jurnal Diferensial*, vol. 6, no. 2, pp. 170–178, 2024. DOI: 10.35508/jd.v6i2.16278.

[18] J. Qian, Y. Zhou, X. Han, and Y. Wang, “Mdbscan: A multi-density dbscan based on relative density,” *Neurocomputing*, vol. 576, p. 127 329, 2024. DOI: 10.1016/j.neucom.2024.127329.

[19] Y. Liu, X. Li, C. Sun, Q. Dong, Q. Yin, and B. Yan, “An indoor thermal comfort model for group thermal comfort prediction based on k-means++ algorithm,” *Energy and Buildings*, vol. 327, p. 115 000, 2025. DOI: 10.1016/j.enbuild.2024.115000.

[20] P. Sarah, S. Krishnapriya, S. Saladi, Y. Karuna, and D. P. Bavirisetti, “A novel approach to brain tumor detection using k-means++, sgldm, resnet50, and synthetic data augmentation,” *Frontiers in Physiology*, vol. 15, p. 1 342 572, 2024. DOI: 10.3389/fphys.2024.1342572.

[21] A. Putra, D. Abdullah, and M. Daud, “Klasterisasi kualitas biji kopi berdasarkan taraf penyusutan menggunakan metode k-harmonic means dengan validasi silhouette index dan c-index,” *Jurnal Janitra Informatika dan Sistem Informasi*, vol. 4, no. 2, pp. 74–86, 2024. DOI: 10.59395/f1jg3b72.