



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

Zalfa Zaliana Nugrahanto¹ 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 disasters remain one of the most frequent natural hazards in Indonesia, particularly in East Java Province, where variations in rainfall, population density, and land use contribute to differing levels of flood vulnerability. Although numerous studies have explored flood susceptibility using geospatial and statistical models, comparative analyses of clustering algorithms specifically designed for complex regional topographies remain limited. This study aims to implement and compare two clustering algorithms Density Based Spatial Clustering of Applications with Noise (DBSCAN) and K-Means++ to map flood vulnerability patterns across the region. The urgency of this study lies in the province's increasing flood frequency and its impact on infrastructure and livelihoods, which demand an adaptive and data driven spatial analysis approach. DBSCAN was selected for its ability to detect irregular, non-linear cluster shapes, while K-Means++ offers efficiency and stability through improved centroid initialization. The dataset includes hydrometeorological and socio-environmental indicators such as rainfall, elevation, slope, land cover, and population density. Clustering performance was assessed using the Silhouette Index (SI) and Davies Bouldin Index (DBI). Results showed that DBSCAN achieved a higher SI (0.3266) compared to K-Means++ (0.2453), indicating better cohesion and separation. Spatially, DBSCAN generated four distinct clusters corresponding to actual flood prone areas, particularly in Jember, Lumajang, Pasuruan, and Sidoarjo. These findings suggest that density-based clustering provides a more reliable representation of heterogeneous spatial flood patterns, supporting local governments in targeted mitigation planning and regional disaster risk management.

Keywords: Flood vulnerability, DBSCAN, K-Means++, Clustering, East Java Province, Silhouette Index, Davies–Bouldin Index, Spatial analysis

Copyright © 2025 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (<https://creativecommons.org/licenses/by-sa/4.0>)

1 Introduction

Flood hazard evaluation has undergone rapid transformation through the use of spatial analytics and machine learning techniques. Recent research confirms that clustering algorithms are powerful for detecting areas vulnerable to flooding by examining multiple environmental indicators simultaneously [1], [2]. For example, in Jakarta, the HDBSCAN approach combined with ArcGIS Pro was utilized to construct an early warning model by examining variables such as rainfall,

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

drainage networks, and population concentration [3]. This method successfully grouped flood points based on spatial density, offering new insights for disaster preparedness and illustrating the practical value of unsupervised learning in flood management [4].

Among clustering techniques, density based approaches have shown remarkable capability in representing complex spatial distributions [5]. DBSCAN, for instance, has been adapted with the FlowSort method to evaluate regional flood hazards by linking hazard, exposure, and vulnerability components [6]. This combination increased the consistency and automation of flood risk assessments and effectively detected regions of critical concern. In other applications, DBSCAN integrated with Kernel PCA achieved high performance in evaluating groundwater quality, producing optimal results when parameters were tuned at $\varepsilon = 0.05$ and $\text{MinPts} = 3$, as confirmed by the Silhouette and Davies Bouldin indices [7].

The K-Means family of algorithms has also contributed greatly to flood susceptibility studies [8]. In Imo State, Nigeria, researchers implemented a geospatial K-Means approach considering ten environmental variables such as slope, elevation, and drainage density [4]. Their findings indicated a silhouette index of 0.56, revealing that about 38% of the region faced a high risk of flooding. The approach effectively delineated vulnerable zones across the landscape without depending on administrative boundaries, providing valuable information for regional land use planning [9].

Comparative analyses between clustering models reveal that each algorithm demonstrates unique strengths. In industrial process monitoring, evaluation of over 300,000 observations from distillation systems showed that DBSCAN, hierarchical, and K-Means methods can identify operational anomalies in real time [10]. In digital behavior studies, K-Means exhibited a balanced compromise between cluster quality and computation time, as confirmed by indices such as Silhouette, Davies Bouldin, and Calinski Harabasz [11].

Further improvements to clustering algorithms have emerged through hybrid and metaheuristic combinations [12]. The integration of K-Means with the Ant Lion Optimization framework enhanced clustering compactness and F measure accuracy relative to single algorithms such as DBSCAN [8]. Likewise, the K-Means++ model employed for risk assessment in the rice supply chain achieved about 93% accuracy, surpassing earlier machine learning approaches including back propagation and radial basis function methods [13]. These developments underline the ongoing refinement of clustering methods across scientific domains.

The scope of clustering applications extends beyond hydrological analysis. For instance, DBSCAN and ST DBSCAN have been applied to cultural heritage mapping to uncover spatial temporal patterns within traditional Chinese settlements [14]. Meanwhile, research on carbon monitoring used DBSCAN for efficient data grouping, effectively minimizing redundancy and lowering mean square error in dense datasets [5].

Despite widespread application, there remains a scarcity of comparative studies focusing specifically on DBSCAN and K-Means++ in the context of flood vulnerability in complex geographical settings such as East Java. Prior works mostly focused on applying a single algorithm, DBSCAN for flood risk classification [6] or K-Means for susceptibility mapping in regions like Nigeria and China [4], [15], while studies comparing their relative strengths are limited. Although research on urban flood resilience has expanded [1], direct algorithmic comparisons in heterogeneous topographical regions are still lacking, leaving an important research gap.

East Java, a province with frequent and severe flooding, is characterized by intense rainfall, complex terrain, and high population density spanning mountains, lowlands, and coastal zones. These diverse conditions call for analytical methods capable of identifying fine spatial variations in flood vulnerability [16]. Given the increasing frequency and impact of floods, data oriented spatial modeling is urgently needed to support more effective mitigation and planning policies across the province [17], [18].

Building upon methodological insights from groundwater monitoring [7], industrial anomaly detection [10], behavioral pattern exploration [11], and previous flood susceptibility mapping

studies [4], this research systematically evaluates and compares the performance of DBSCAN and K-Means++ using multiple internal validation metrics to determine the most appropriate model for identifying flood vulnerability patterns in East Java Province.

2 Methods

2.1 Research Procedure

The research was conducted systematically through several main stages as follows:

1. **Data Input and Standardization**

The collected dataset was first entered into the analysis system. Each variable was standardized using the *Z-score* transformation to ensure consistency of measurement scales and to prevent bias among indicators [4].

2. **Multicollinearity Test**

The test was performed by examining the *Variance Inflation Factor* (VIF) and *Tolerance* values to ensure no high linear correlations existed among independent variables[6].

3. **Clustering Analysis Using DBSCAN**

The DBSCAN method was applied by calculating the distance between data points using the *Euclidean Distance*, determining parameters ε and *MinPts* through the *K-Nearest Neighbors* (KNN) and *Elbow* approaches[7], followed by cluster formation and spatial visualization using QGIS[3].

4. **Clustering Analysis Using K-Means++**

The number of clusters (K) was selected according to the study objectives. Cluster initialization was performed with the K-Means++ method, followed by iterative centroid recalculations until convergence was achieved, producing stable and distinct groupings.

5. **Evaluation and Interpretation**

The validity of clustering results from both methods was evaluated using the *Silhouette Index* (SI) and the *Davies–Bouldin Index* (DBI). The final clusters were visualized and interpreted to identify the spatial patterns of flood-prone areas.

2.2 Data Standardization

Before performing the clustering analysis, all research variables were standardized to eliminate the effect of different measurement scales [19]. 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 [20].

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

$$Z_{ij} = \frac{X_{ij} - \bar{X}_j}{s_j} \quad (1)$$

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 Multicollinearity Test

Prior to clustering, a multicollinearity test was conducted to ensure that no strong linear relationships existed among the independent variables. High multicollinearity can distort the

distance calculation and reduce the reliability of clustering outcomes [21].

The test was performed using two indicators: the *Variance Inflation Factor* (VIF) and the *Tolerance* value. The VIF for each variable is calculated as follows [22]:

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad (2)$$

where R_j^2 represents the coefficient of determination obtained when variable j is regressed against all other independent variables [23]. The *Tolerance* value is defined as:

$$\text{Tolerance}_j = 1 - R_j^2 \quad (3)$$

A variable is considered to exhibit multicollinearity if $\text{VIF} > 10$ or $\text{Tolerance} < 0.10$ (Hair et al., 2019). Variables exceeding these thresholds were evaluated and, if necessary, excluded from further analysis to maintain model stability.

This procedure ensured that all variables used in the DBSCAN and K-Means++ analyses were statistically independent and contributed uniquely to the clustering process.

2.4 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 [24]. DBSCAN is particularly effective for identifying arbitrary-shaped clusters and distinguishing noise or outlier data [25].

DBSCAN requires two key parameters: the neighborhood radius (ε) and the minimum number of points (*MinPts*) required to form a dense region [26]. The distance between two data points was measured using the *Euclidean Distance*, formulated as:

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

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:

$$N_\varepsilon(x_i) = \{x_j \in D \mid d(x_i, x_j) \leq \varepsilon\} \quad (5)$$

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 optimized using the *K-Nearest Neighbors* (KNN) and *Elbow* methods to identify suitable density thresholds. The resulting clusters were then visualized spatially using QGIS to observe the distribution and pattern of flood-prone areas.

2.5 K-Means++

The K-Means++ algorithm was employed to perform partition-based clustering by minimizing the total variance within clusters [27]. 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 [28]. The probability for selecting a data point x_i as the next centroid is defined as:

$$P(x_i) = \frac{D(x_i)^2}{\sum_{k=1}^n D(x_k)^2} \quad (6)$$

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*:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^p (x_{ik} - c_{jk})^2} \quad (7)$$

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

$$c_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i \quad (8)$$

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 flood-prone regions.

2.6 Cluster Validation and Evaluation

The performance and reliability of the clustering outcomes produced by both DBSCAN and K-Means++ algorithms were examined using two internal validation indices: the *Silhouette Index (SI)* and the *Davies–Bouldin Index (DBI)*. These indices quantitatively measure how compact the clusters are internally and how well they are separated from one another, providing an objective evaluation of clustering effectiveness.

2.6.1 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 [29]:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (9)$$

where:

- $a(i)$: the average distance between observation i and all other points within the same cluster,
- $b(i)$: 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:

$$SI = \frac{1}{n} \sum_{i=1}^n s(i) \quad (10)$$

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.

2.6.2 Davies-Bouldin Index (DBI)

The Davies–Bouldin Index measures the average similarity between each cluster and its most similar counterpart, where similarity is defined as the ratio of within-cluster scatter to between-cluster separation. The DBI is formulated as [30]:

$$DBI = \frac{1}{K} \sum_{i=1}^K \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (11)$$

where:

- $a(i)$: the average distance between observation i and all other points in the same cluster,
- $b(i)$: the minimum average distance between observation i and all points in other clusters.

A lower DBI value indicates more compact and well-separated clusters, while higher values suggest overlapping or less distinct groupings.

In this study, both indices were used to compare the performance of DBSCAN and K-Means++ in identifying spatial patterns of flood-prone areas. The combination of SI and DBI provided a robust evaluation framework for assessing the effectiveness and reliability of the clustering results.

3 Results and Discussion

3.1 Evaluation of DBSCAN Parameters and Cluster Formation

Before running the DBSCAN algorithm, it is necessary to determine the optimal values of the two main parameters, namely *Epsilon* (ε) and *Minimum Points* (*MinPts*). 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 $\epsilon = 0.8$ and $MinPts = 3$, which produced a total of four (4) distinct clusters and eleven (11) 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 areas with high flood intensity and significant infrastructure impact; Cluster 1 represents areas with moderate to low flood occurrences; while Clusters 2–4 represent specific patterns influenced by geographical and hydrological conditions, such as midland flooding or tidal inundation.

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

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

The results demonstrate that DBSCAN successfully identified spatially coherent groupings of flood-prone regions across East Java Province. These clusters can be further visualized on a spatial map to assist in flood risk management and policy planning at the local government level.

Based on the results of the DBSCAN clustering analysis, the spatial distribution of flood-prone regions in East Java Province can be visualized as shown in Figure 1. This map illustrates the division of districts and cities into several clusters according to their flood characteristics, which were formed based on variables such as the number of flood incidents, affected population, and infrastructure damage. The visualization integrates the results of the DBSCAN and K-Means++ algorithms to present a comparative spatial perspective of regional flood vulnerability.

Each color in the map corresponds to a cluster group, namely: Cluster 0 (green), Cluster 1



Figure 1: DBSCAN

(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.

As illustrated in Figure 1, the cluster distribution shows a clear spatial pattern across East Java. Regions classified into **Cluster 0** (green) are those with the highest frequency and severity of floods, including Jember, Lumajang, Pasuruan, and Sidoarjo. **Cluster 1** (orange) encompasses districts with moderate flood intensity and limited infrastructure impact, such as Kediri, Trenggalek, and Ngawi. **Cluster 2** and **Cluster 3** represent areas with medium-level characteristics, typically in the central and western regions of the province, including Bondowoso, Bojonegoro, and Kota Pasuruan. Meanwhile, **Cluster 4** (light green) primarily covers the northern coastal zones such as Gresik and Lamongan, which are influenced by tidal inundation and low-lying topography.

This spatial pattern suggests that flood vulnerability in East Java is influenced by both topographical and hydrometeorological factors. Southern and eastern regions tend to have higher rainfall and steeper terrain, resulting in more severe flood events, while northern coastal regions experience more frequent but lower-intensity inundations. These findings highlight the importance of cluster-based spatial analysis as a basis for developing targeted flood mitigation strategies and regional disaster preparedness planning.

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.

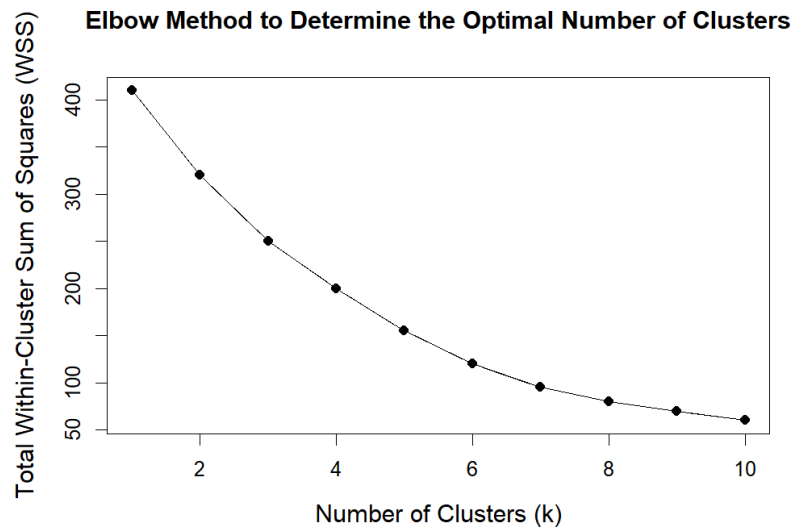


Figure 2: Elbow

As illustrated in Figure 2, the inflection point occurs at $k = 3$, indicating that three clusters provide the best trade-off between compactness and separation. Hence, the number of clusters for this study was set to three.

3.2.2 Cluster Formation Using K-Means++

After determining the optimal number of clusters, the K-Means++ algorithm was executed with the following parameters: `num_init = 10`, `max_iters = 100`, and `initializer = "kmeans++"`. Each observation (district or city) was assigned to the cluster whose centroid was closest based on Euclidean distance. The algorithm iteratively minimized the total within-cluster variance until convergence was achieved.

1. Cluster 1 consists of 34 districts/cities with moderate flood characteristics.
2. Cluster 2 consists of 1 district/city with localized but high flood impact.
3. Cluster 3 consists of 3 districts/cities with high flood frequency and severe damage intensity.

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++)

Cluster	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	2.588	0.118	0.029	5096.06	0.176	0.000	2.000	895.47	0.059	0.118	0.000
2	2.000	2.000	0.000	6405.00	10.000	10.000	16.000	1431.00	0.000	0.000	0.000
3	10.333	0.667	1.000	48160.67	0.000	0.000	0.000	7603.00	0.333	0.000	0.333

Based on Table 3, Cluster 1 represents the majority of districts with moderate flood frequencies and damage levels, as reflected by moderate values in **X1** (number of flood events) and **X4** (affected population). Cluster 2, consisting of only one region, exhibits higher values in **X5**, **X6**, and **X7**, suggesting a localized area that experienced severe flood impact. Cluster 3, with three regions, has extremely high values in **X4** and **X8**, indicating severe and widespread flood events.

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 ($\uparrow$)
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 flood-prone regions 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 their flood risk intensity and impact indicators.

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 following interpretations can be made:

1. **Cluster 0** represents areas with the *highest average flood impact*, marked by dominant values in X4 (total affected population) and X8 (total damage). These are the regions that consistently face the most severe flood consequences, where inundation events tend to affect both livelihoods and infrastructure on a large scale. Such conditions are commonly observed in low-lying settlements along major river basins, where dense population and poor drainage exacerbate the impact of extreme rainfall events [17].
2. **Cluster 1** consists of districts with relatively low averages across all indicators, pointing to areas that only experience **light or infrequent flooding**. These regions generally enjoy higher elevation, better watershed management, and greater soil permeability—factors that naturally limit flood intensity. For instance, northern Jombang displays this kind of resilience, with its karstic geological layers allowing faster water infiltration and shorter flood duration [18].
3. **Cluster 2** displays very low values for most variables yet shows a mild anomaly in X10, suggesting small-scale or episodic flood events. These cases likely occur in confined catchment areas or semi-urban pockets where localized rainfall extremes briefly overwhelm drainage systems before conditions stabilize. This pattern mirrors the scattered, isolated flood hotspots seen in East Java’s province-wide hazard mapping, where risk intensity fluctuates sharply between neighboring subdistricts [17].

4. **Cluster 3** exhibits high values in X1 (flood frequency) and X4 (affected population), reflecting **frequent but spatially concentrated flooding**. This behavior is characteristic of densely populated districts that experience recurring waterlogging during heavy rain seasons—often near rivers or low-lying drainage basins. The Brantas River corridor is a classic example of this, as many communities in its vicinity report repeated inundation events every rainy season despite relatively contained spatial spread [18].
5. **Cluster 4** shows consistently elevated values in X1, X4, and X8, capturing areas with **moderate to severe flood exposure** that often sit between urban expansion zones and agricultural lowlands. These semi-urban landscapes tend to accumulate surface runoff during prolonged rainfall, creating chronic flooding patterns that disrupt daily activities and damage property. Studies across East Java also emphasize that such mid-level flood zones are the most widespread, reflecting the province’s transitional topography and mixed land-use structure [17], [18].

Overall, the DBSCAN clustering effectively separated the districts and cities in East Java into five distinct groups based on their flood characteristics. Clusters 0 and 3 clearly stand out as the most critical regions requiring prioritized disaster mitigation and resource allocation. Meanwhile, Clusters 1 and 2 represent low-risk areas with minimal flood occurrences, while Cluster 4 serves as a transitional category between moderate and severe flood conditions.

4 Conclusion

This research compared the performance of DBSCAN and K-Means++ algorithms in clustering flood vulnerability across East Java Province. The results indicated that DBSCAN achieved a higher Silhouette Index value of 0.3266 compared to K-Means++ with 0.2453, reflecting stronger cohesion and clearer separation between clusters. The density-based approach also succeeded in identifying spatially coherent high-risk zones, particularly in Jember, Lumajang, Pasuruan, and Sidoarjo, which correspond to areas historically affected by flooding events.

These findings highlight that DBSCAN is more effective for analyzing complex hydrometeorological data and heterogeneous topographical conditions, while K-Means++ performs better on relatively uniform datasets. The study confirms that density-based clustering methods can enhance regional flood mapping accuracy and provide valuable insights for data-driven disaster mitigation and planning.

Despite its promising results, this study has limitations related to dataset size and the use of a single-year observation period. Future research can expand the analysis by integrating temporal flood data, exploring hybrid clustering models, and incorporating additional environmental parameters to improve accuracy and robustness. Overall, this work contributes to developing a more adaptive and evidence-based approach to flood risk management in East Java Province.

CRedit Authorship Contribution Statement

Clearly define each author’s contributions using the Contributor Roles Taxonomy (CRedit). Indicate contributions using standardized terms such as Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing–Original Draft Preparation, Writing–Review & Editing, Visualization, Supervision, Project Administration, and Funding Acquisition.

For example: **Author One:** Conceptualization, Methodology, Writing–Original Draft. **Author Two:** Data Curation, Formal Analysis, Writing–Review & Editing. **Author Three:** Software, Validation, Visualization. **Author Four:** Supervision, Project Administration, Funding Acquisition.

Declaration of Generative AI and AI-assisted technologies

Authors must explicitly disclose if generative AI or AI-assisted technologies were used during the preparation of the manuscript. If such tools were utilized, clearly state their role and purpose, for example, in data analysis, writing assistance, proofreading, or generating graphical outputs. Provide the names and versions of the specific tools (e.g., ChatGPT version 4, Grammarly, etc.). If no AI or AI-assisted technologies were involved in the preparation of your manuscript, state clearly: “No generative AI or AI-assisted technologies were used during the preparation of this manuscript.”

Declaration of Competing Interest

Authors must declare any potential competing or conflicting interests that could be perceived as influencing the outcomes or interpretation of their research. Conflicts of interest include financial or personal relationships with people or organizations that could inappropriately affect the research outcomes. Clearly state: “The authors declare no competing interests,” or explicitly describe the nature of any competing interests if applicable.

For example: “The authors declare no competing interests,” or “Author A has received consulting fees from Company X. Authors B and C declare no competing interests.”

Funding and Acknowledgments

This section must acknowledge all sources of financial support received during the research and manuscript preparation. Clearly state the names of funding organizations, grant numbers, or project codes as appropriate. If there was no external financial support, explicitly state: “This research received no external funding.”

Additionally, authors can use this section to acknowledge individuals or institutions who provided significant assistance, facilities, materials, or guidance that contributed substantially to the completion of the research but did not fulfill authorship criteria.

Data and Code Availability

In this section, authors must clearly state where and how the research data and code supporting the findings of this study can be accessed. Provide specific details regarding data repositories, source code platforms (e.g., GitHub, GitLab), accession numbers, URLs, or DOI links. If the data or code are available upon reasonable request, clearly specify the procedure for accessing them, including the contact details of the corresponding author. If data or code cannot be shared openly due to ethical, legal, or privacy reasons, state explicitly: “The data and code supporting the findings of this study are available from the corresponding author upon reasonable request and subject to confidentiality agreements.”

For example: “The dataset and code analyzed during the current study are publicly available in the Kaggle repository¹ and GitHub².”

¹<https://www.kaggle.com/datasets/hendratno/covid19-indonesia>

²<https://github.com/username/project-repo>

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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1016/j.jafrearsci.2025.105872).
- [5] R. Ma, J. Sha, S. Zhang, D. Zhu, W. Kang, and J. Liu, “Fast grouping fusion method of dual carbon monitoring data based on DBSCAN clustering algorithm,” *Results in Engineering*, vol. 26, p. 105 057, 2025. DOI: [10.1016/j.rineng.2025.105057](https://doi.org/10.1016/j.rineng.2025.105057).
- [6] 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](https://doi.org/10.1016/j.jhydrol.2024.131587).
- [7] 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](https://doi.org/10.1016/j.jhydrol.2025.133566).
- [8] 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](https://doi.org/10.1016/j.kijoms.2018.09.001).
- [9] G. Vardakas, I. Papakostas, and A. Likas, “Deep clustering using the soft silhouette score: Towards compact and well-separated clusters,” *arXiv preprint arXiv:2402.00608*, 2024.
- [10] 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](https://doi.org/10.1016/j.jii.2025.100970).
- [11] 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](https://doi.org/10.1016/j.ins.2025.122745).
- [12] L. Ding, C. Li, D. Jin, and S. Ding, “Survey of spectral clustering based on graph theory,” *Pattern Recognition*, vol. 151, p. 110 366, 2024.
- [13] B. Qi et al., “A risk assessment model for the entire rice processing chain based on K-Means++ and extreme learning machine,” *LWT*, vol. 223, p. 117 803, 2025. DOI: [10.1016/j.lwt.2025.117803](https://doi.org/10.1016/j.lwt.2025.117803).
- [14] Y. Kaijian, Y. Congjian, and S. Jun, “Agglomeration phenomena and evolutionary patterns of contemporary architectural interventions in traditional villages: An empirical analysis from 211 cases in songyang county, china,” *Habitat International*, vol. 166, p. 103 571, 2025. DOI: [10.1016/j.habitatint.2025.103571](https://doi.org/10.1016/j.habitatint.2025.103571).
- [15] 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](https://doi.org/10.1016/j.jhydrol.2018.06.060).

- [16] E. Hidayah, S. Husrin, Y. Prasetyo, D. Rahmawati, and F. Nugraha, "Assessing coastal flood susceptibility in east java, indonesia," *Water*, vol. 14, no. 23, p. 3869, 2022. DOI: [10.3390/w14233869](https://doi.org/10.3390/w14233869). Available online.
- [17] 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. 012081. DOI: [10.1088/1742-6596/1825/1/012081](https://doi.org/10.1088/1742-6596/1825/1/012081). Available online.
- [18] 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](https://doi.org/10.12962/j27745449.v2i3.100). Available online.
- [19] 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.
- [20] 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. 105727, 2025. DOI: [10.1016/j.actpsy.2025.105727](https://doi.org/10.1016/j.actpsy.2025.105727).
- [21] E. Yaldi et al., "Penerapan uji multikolinieritas dalam penelitian manajemen sumber daya manusia," *Jurnal Ilmiah Manajemen Dan Kewirausahaan (JUMANAGE)*, vol. 1, no. 2, pp. 94–102, 2022.
- [22] F. Nadya, "Estimator modified jackknife untuk mengatasi multikolinieritas pada regresi poisson (studi kasus: Angka kematian bayi di provinsi sumatera utara)," Ph.D. dissertation, UNIMED, 2024.
- [23] M. U. Abshor, L. B. Hasiholan, and D. Malik, "Pengaruh harga, promosi dan kualitas produk terhadap keputusan pembelian produk torabika duo (studi kasus di area kabupaten demak)," *Journal of Management*, vol. 4, no. 4, 2018.
- [24] 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.
- [25] J. Qian, Y. Zhou, X. Han, and Y. Wang, "Mdbscan: A multi-density dbscan based on relative density," *Neurocomputing*, vol. 576, p. 127329, 2024.
- [26] 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.
- [27] 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. 115000, 2025.
- [28] 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. 1342572, 2024.
- [29] 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.
- [30] Y. Hasan, "Pengukuran silhouette score dan davies-bouldin index pada hasil cluster k-means dan dbscan," *KAKIFIKOM (Kumpulan Artikel Karya Ilmiah Fakultas Ilmu Komputer)*, pp. 60–74, 2024.