PYTHAGORAS: Jurnal Program Studi Pendidikan Matematika, 14(1): 111-122
April 2025
p-ISSN: 2301-5314
e-ISSN: 2615-7926

Density based spatial clustering of application with noise using flower pollination algorithm for leptospirosis clustering

Finansiya S. Abd. Karim[1], Emli Rahmi[2], Siti Nurmardia Abdussamad[3,*], Isran K. Hasan[4], Nisky Imansyah Yahya[5]

[1,3,4]Department of Statistics, Faculty of Mathematics and Science, Gorontalo State University
[2,5]Department of Mathematics, Faculty of Mathematics and Science, Gorontalo State University
[*]e-mail: sitinurmardia@ung.ac.id

Diserahkan: 13/02/2025; Diterima: 05/05/2025; Diterbitkan: 08/05/2025

Abstract. Leptospirosis is an important health problem in Indonesia, with most cases found in East Java and Central Java provinces. This study aims to identify the distribution pattern of leptospirosis in the two provinces using a clustering approach. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method is used to cluster areas based on leptospirosis spread factors, but DBSCAN requires optimal parameter determination for accurate results. Therefore, this research implements Flower Pollination Algorithm (FPA) to optimize the epsilon (ϵ) and minimum points (MinPts) parameters in DBSCAN. This research uses secondary data obtained from data on the Number of Natural Disaster Events by Regency / City in East Java and Central Java Provinces in 2023 and data on Population Density by Regency / City in East Java and Central Java Provinces in 2023. The population in this study uses all observations, namely all people in the districts and cities in East Java and Central Java. The sampling technique is saturated sampling, that is, the entire population in the study is sampled. The clustering results using FPA-DBSCAN resulted in two main clusters, with 30 districts/municipalities detected as noise, 23 districts/municipalities belonging to cluster 0, and 20 districts/municipalities in cluster 1. The validation test using Silhouette Coefficient showed a value of 0.1892, indicating that the clustering is quite valid. The results of this clustering can serve as a strategic reference for local governments in optimizing disease surveillance and targeted health interventions.

Keywords: Clustering, Density Based Spatial Clustering of Application with Noise, Flower Pollination Algorithm, Leptospirosis

Introduction

Infectious diseases are one of the leading causes of death in Indonesia, with leptospirosis being one of the most prominent (Achjar et al., 2024). Leptospirosis is an infectious disease transmitted through direct contact with animals or exposure to an environment contaminated with animal urine (Purnama & Hartono, 2022). The disease has varied symptoms, ranging from mild symptoms to serious complications such as kidney failure and pulmonary hemorrhage that can cause breathing difficulties (Aziz & Suwandi, 2019). Based on data from the International Leptospirosis Society (ILS) in 2011, Indonesia ranked third in the world for leptospirosis incidence. Data from the Indonesian Ministry of Health in 2023 recorded fluctuations in the number of leptospirosis cases, where in 2019 there were 921 cases and increased to 1,170 cases in 2020. This figure then dropped to 736 cases in 2021, but again increased to 1,624 cases in 2022, and reached its peak in 2023 with 2,554 cases, with a mortality rate (CFR) of 8%. East Java and Central Java provinces accounted for the highest number of cases, 42.7% and 36.6% of the total national cases respectively.

Environmental risk factors that have the potential to cause leptospirosis include a history of flooding (Zukhruf & Sukendra, 2020), high risk in livestock breeders and farmers (Dewi & Yudhastuti, 2019), bamboo house wall conditions (Afiff, Adi, Saraswati, & Wuryanto, 2019), population density (Aziz & Suwandi, 2019), and proximity of houses to sewers or waste that become rat habitat (Purnama & Hartono, 2022). Given the importance of these factors, this study aims to identify patterns of leptospirosis disease spread using a clustering approach, which can help the government design more targeted prevention policies.

Clustering is a data grouping technique used to classify objects or data patterns into several relatively similar clusters, where similar objects or data patterns are placed in the same cluster, while different objects or data patterns are separated into different clusters (Abdussamad, Astutik, & Effendi, 2020). One clustering method that is often used is Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which excels in detecting clusters with irregular shapes and is able to handle noise (Id, Astrid, & Mahdiyah, 2017). However, DBSCAN has shortcomings in determining optimal parameters, such as the input values of ϵ (maximum distance between points) and MinPts (minimum number of data points required to form a cluster). Improper parameter determination can result in less accurate clusters. To overcome this weakness, this research proposes the use of Flower Pollination Algorithm (FPA) to optimize the parameters epsilon and minimum points in DBSCAN. FPA is an optimization algorithm inspired by the process of flower pollination (Ali, Siswanto, & Baehaqi, 2024), and is proven to be more efficient in solving non-linear optimization problems compared to other algorithms such as Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) (Yang, 2012).

Previous research focusing on analysing the factors that cause leptospirosis has been conducted. Research by Ariani & Wahyono (2020) used Binary Logistic Regression method to analyse factors affecting the incidence of leptospirosis. Their results showed that a history of flooding and contact with sewers were factors that influenced the incidence of leptospirosis, while the dominant factor was the presence of rats. Husni, Martini, Suhartono, Budiyono, & Raharjo (2023) used Chi-Square and Logistic Regression methods in their study, and found that garbage and waste disposal conditions influenced the incidence of leptospirosis. Dewi & Yudhastuti (2019) also used the Chi-Square method and found that the presence of sewers and the use of personal protective equipment (PPE) influenced the incidence of leptospirosis. However, these studies did not address the aspect of clustering areas based on the characteristics of leptospirosis spread factors, which is important for designing more appropriate and effective health policies.

Therefore, this study aims to evaluate the clustering results using FPA-DBSCAN and cluster the factors that cause the spread of leptospirosis in East Java and Central Java. By using this method, it is hoped that more precise clustering can be obtained in understanding the pattern of disease spread, which in turn can help the government in designing more effective and efficient health policies to reduce the increasing number of leptospirosis sufferers in East Java and Central Java.

Research Methods

This study uses secondary data obtained from the official websites of BPS of East Java and Central Java Provinces (*jatim.bps.go.id* and *jateng.bps.go.id*). The data includes six variables

selected based on the literature that has identified the relationship of these variables with the risk of leptospirosis spread. The variables used were:

1. X_1 : Number of flood events (frequency)
2. X_2 : Population density (people/km^2)
3. X_3 : Percentage of houses with bamboo or wicker structures (%)
4. X_4 : Percentage of houses with distance from sewer or sewage < 10 metres (%)
5. X_5 : Number of farmers (person/worker)
6. X_6 : Number of breeders (person/worker)

Interactions between variables in the context of the spread of leptospirosis show complex relationships that reinforce each other. High population density often correlates with sanitation challenges, and this is exacerbated by the proximity of houses to sewers and the high proportion of bamboo-walled houses, which together create an ideal environment for rat populations as the main vector of leptospira. On the other hand, the high number of farmers and herders reflects the vulnerability of the community due to the intensity of their interactions with wet environments and livestock, which are potential sources of leptospira bacterial contamination. In addition, flooding is a contributing factor as it expands the area of water contamination and increases people's exposure to polluted environments. Thus, the combination of all these variables forms a significant cumulative risk for the spread of leptospirosis in an area.

Sampling in this study was carried out using the saturated sampling method, which is a sampling technique carried out on all members of the population, which means that all members of the population are samples (Amin, Garancang, & Abunawas, 2023). This technique was chosen because the available data coverage includes the entire administrative population, as well as to ensure that no information is missed in spatially mapping the risk of leptospirosis.

The clustering method used in this research is Density Based Spatial Clustering of Application with Noise (DBSCAN) with Flower Pollination Algorithm (FPA) optimisation. FPA is an optimisation algorithm inspired by the flower pollination process, where pollen is moved between flowers by insects, wind, or other mechanisms to produce new flowers (Ali et al., 2024). The distance calculation in the DBSCAN algorithm uses the Euclidean distance method. The DBSCAN algorithm uses euclidean distance to calculate the close character between a data point and a predetermined group centre point (Putri et al., 2021). This research uses K-Nearest Neighbor as the generation of the initial value of the parameters of the minimum points and the Elbow method in determining the initial value of epsilon. The index test carried out to evaluate the clustering results is using the silhouette coefficient.

The stages of the research conducted are described as follows:

1. Data input.
2. Descriptive analysis of the variables used.
3. Conducting multicollinearity test. The multicollinearity test includes an assumption test to ensure that there is no correlation between the variables. One way to test the

multicollinearity assumption is to use the Variance Inflation Factor (VIF). The VIF value is calculated using the following formula (Sriningsih, Hatidja, & Prang, 2018):

$$VIF = \frac{1}{1 - R^2} \quad (1)$$

with R^2 is the coefficient of determination between independent variables. If the VIF value> 10 means there is a multicollinearity problem, if the VIF value ≤ 10 means there is no multicollinearity problem (Ramadani, 2021). If there is multicollinearity then repeat the first stage.

4. Standardising the data. Data standardisation is the process of transforming data so that it has a distribution with a mean of zero and a standard deviation of one (Elfaladonna, Sartika, & Putra, 2024). Data standardisation aims to reduce unit differences between variables and equalise variable scales (Warolemba, Resmawan, & Isa, 2023). As the data in this study had different units, standardisation was applied by converting the data to z-scores. The standardisation process with z-score uses the following formula (Puspita, 2021):

$$Z = \frac{x_i - \bar{x}}{\sigma} \quad (2)$$

Description:

Z : Standardised (z-score) of a variable.
x : The data value of a variable.
$\bar{x}$: The average data (mean) of a variable.
σ : Standard deviation of a variable.

5. Perform distance calculations using the euclidean distance with the following formula (Pribadi, Yunus, & Wiguna, 2022):

$$D_{(x,y)} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

Description:

$D_{(x,y)}$: The euclidean distance between point x and point y.
i : Data for each observation.
x_i : The centre data of the-*i* cluster.
y_i : The data of the-*i* attribute.
n : The amount of data.

6. Perform parameter tuning using the K-Nearest Neighbors (KNN) method, which groups data based on its closest distance to other data or neighbours, using the formula in equation (3). The purpose of KNN is to generate the initial value of the MinPts parameter. In addition, parameter tuning is performed to generate the initial value of the Eps parameter, using the Elbow method. This method is a visual technique to determine the optimal number of clusters by comparing the Sum of Square Error (SSE) values at each number of clusters. A sharp decrease in SSE between the first and second number of clusters indicates the optimal point which is visible as an arc on the graph (Samosir, Widodo, Anwar, Sekti, & Erzed, 2024). The formula for the elbow method is as follows (Samosir et al., 2024):

$$SSE = \sum_{k=1}^{k} \sum_{x_i \epsilon S_k} ||\boldsymbol{X}_i - C_k||^2 \quad (4)$$

Description:

k : Number of cluster.

x_i : The attribute value of the-*i* data.

C_k : Number of clusters *i* in the-*k* cluster.

|| : Calculation of euclidean distance.

Parameter tuning using these two methods was performed resulting in values of minimum points = 13 and epsilon = 1.9.

7. Optimising the minimum points and epsilon parameter values using the Flower Pollination Algorithm (FPA). FPA is an optimisation algorithm inspired by the process of flower pollination in nature, where pollen from one flower is transferred to another by insects, wind, or other mechanisms, to produce a new flower (Ali et al., 2024). The creatures that assist in this pollination process are called pollinators. The selected flower is the one with the best appearance, based on this, the FPA method can be applied (Munir, Wartana, & Suslistiowati, 2023), with steps:
   a) Enter the FPA parameters of flower population = 20 and probability switch = 0.8 (Sakti & Putra, 2019). Dimension 6, the value of epsilon = 1.9 and minimum points = 13 as the initial generation value.
   b) Initialise the initial population with a random solution.
   c) Finding the temporary best solution from the initial population.
   d) Determine global pollination or local pollination by looking at the random walk and probability switch values. Probability switch p ϵ [0,1] is a parameter that regulates the occurrence of global pollination or local pollination, where ϵ is a random walk. Random walk is a random process consisting of a series of random steps with a random distribution. If the value of random walk < probability switch then perform global pollination with the following equation (Munir et al., 2023):

$$x_i^{t+1} = x_i^t + \gamma L(\lambda)(g_* - x_i^t) \quad (5)$$

   Description:

   x_i^t : The-*i* solution of the-*t* iteration.

   g_* : The most optimal interim solution from each *t* iteration.

   L : Pollinator step size that follows a Lèvy distribution.

   γ : The unit that controls Lèvy's movements.

   λ : The scale that governs Lèvy's movements.

   If not, perform local pollination using the following equation (Munir et al., 2023):

$$x_i^{(t+1)} = x_i^t + \epsilon\left(x_j^t - x_k^t\right) \quad (6)$$

   where x_j^t is pollen taken from flower *j* on the same plant, x_k^t is pollen taken from flower *k* on the same plant, and ϵ is a random number distributed in U(0,1).
   e) Evaluate the provisional best solution with the fitness values of all newly generated solutions.

f) View the maximum iteration limit. If the number of iterations has reached the set maximum limit, the algorithm will stop automatically.
g) Display the best solution.

8. Conducting index testing on the optimal epsilon and minimum points values. Determination of the epsilon and minimum points values is based on the validation results of the Silhoette Coefficient index with the following formula (Harjanto, Vatresia, & Faurina, 2021):

$$S_i = \frac{(b_i - a_i)}{\max(b_i - a_i)} \quad (7)$$

where a_i is the average distance of object i to all objects and b_i is the smallest average value of object *i* to objects in different clusters.

9. Cluster formation by implementing the optimal epsilon and minimum points parameter values in FPA-DBSCAN.
10. Creating mapping according to the clustering results using QGIS software.
11. Interpreting the characteristics of each cluster.

Results and Discussion

Descriptive Statistics

The data used in this study is data on the factors that cause leptospirosis disease which consists of 6 variables previously described. The following are the results of descriptive statistics processed using Python on Jupyter Notebook.

Table 1. Descriptive Statistics

Var	N	Mean	Std	Min	Q1	Median	Q3	Max
X_1	73	2.34	2.43	0.	0.00	2.00	4.00	9.0
X_2	73	2007.16	2315.35	410	744.00	1060.00	1702.00	11302.0
X_3	73	1.27	1.49	0.	0.18	0.90	1.58	7.3
X_4	73	45.72	24.810	0.	23.81	39.03	67.08	91.6
X_5	73	134509.04	89916.47	992	75985.00	141963.00	186837.00	366667.0
X_6	73	78218.52	57888.53	802	23240.00	77817.00	115870.00	218584.0

Table 1 shows notable variability in the six variables observed across East and Central Java districts/cities in 2023. Flood events (X_1) averaged 2.43, population density (X_2) 2,007.16 people/km², and the proportion of bamboo/wicker houses (X_3) 1.27%. The percentage of houses near sewers (X_4) averaged 45.72%, while the number of farmers (X_5) and breeders (X_6) averaged 134,509.04 and 78,218.52, respectively. The table above shows that there is a considerable difference in the range of values between variables.

Multicollinearity Test

A multicollinearity test was conducted to ensure no strong correlations existed among the independent variables, as required for cluster analysis. The following are the results of the multicollinearity test on each variable, which is processed using Python on Jupyter Notebook.

Table 2. Multicollinearity Test

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	-	1.011723	1.006791	1.214404	1.003374	1.003374
X_2	1.011723	-	1.120048	1.011598	2.027151	1.802180
X_3	1.006791	1.120048	-	1.003695	1.273657	1.223442

X_4	1.214404	1.011598	1.003695	-	1.055684	1.076031
X_5	1.003374	2.027151	1.273657	1.055684	-	8.099951
X_6	1.003374	1.802180	1.223442	1.076031	8.099951	-

The multicollinearity test results show that the VIF value between variables is <10, which means there is no multicollinearity. Thus, the assumption of no multicollinearity is met and cluster analysis can proceed.

Data Standardisation

Due to differing scales and units across variables, data standardisation was applied to ensure comparability. The following are the results of data standardisation processed using Python on Jupyter Notebook.

Table 3. Data Standardisation

Index	X_1	X_2	X_3	X_4	X_5	X_6
1	-0.97146547	-0.6945883	-0.11903299	1.60370097	0.0457226	0.71508088
2	-0.97146547	-0.57890798	-0.13254685	1.51432393	0.76826566	1.31302137
3	-0.97146547	-0.61500372	0.27286879	0.89272887	0.46653602	0.88229062
⋮	⋮	⋮	⋮	⋮	⋮	⋮
71	2.76100712	1.13542218	-0.85553808	-0.50495008	-1.37981112	-1.28206054
72	-0.14202712	2.14958198	-0.03794987	-0.39130778	-1.48094386	-1.33326885
73	1.10213042	2.30875114	-0.85553808	0.82033751	-1.4876741	-1.34478376

Distance Calculation

Euclidean distance is used to measure the proximity of data points within an epsilon radius; core points that exceed the minimum threshold and are connected will form a cluster, while the others are classified as noise. The following are the results of distance calculation processed using Python on Jupyter Notebook.

Table 4. Distance Calculation Result Data

No.	1	2	3	4	5	6	7	8	...	73
1.	0.0000	0.9493	0.9321	0.8295	1.5201	1.2843	3.0954	1.4229	...	4.5902
2.	0.9493	0.0000	0.9103	0.9822	0.6634	0.7000	2.2032	1.4756	...	5.0790
3.	0.9330	0.9103	0.0000	0.4645	1.4528	0.8902	2.5904	0.9280	...	4.7860
4.	0.8296	0.9822	0.4645	0.0000	1.4541	0.8134	2.7265	0.9429	...	4.4844
5.	1.5201	0.6634	1.4528	1.4541	0.0000	0.8117	1.7892	1.8682	...	5.3860
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
73.	3.3508	4.0103	3.5894	3.2373	4.3900	3.8228	5.6429	3.4981	...	2.3937

Tunning Parameter

The initial value for minimum points in the FPA algorithm was determined using the K-Nearest Neighbor (KNN) method to explore the optimal parameter search space. The following is the result of generating the minimum initial point value using KNN using Python on Jupyter Notebook.

Table 5. K-Nearest Neighbors Result

Minimum Samples
13

Next, the initial value of epsilon is found using the Elbow method, where the optimal point is determined based on the decrease in inertia that starts to slow down significantly. The point at which this change resembles the shape of an elbow is called the elbow point.

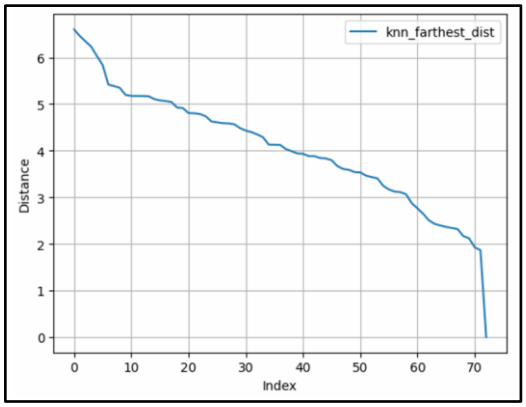

Figure 1. Elbow Chart

Based on Figure 1, it can be seen that the elbow point is formed at point 1.9, which means that the decrease in inertia occurs at that point. So, from this stage, the initial generation values of epsilon and minimum points are found to be 1.9 and 13, respectively. These values will be included as exploration constraints in the FPA optimisation algorithm.

Flower Pollination Algorithm (FPA)

This stage optimised epsilon and minimum points for DBSCAN using FPA, guided by parameter tuning results and considering local or global pollination based on the random walk and probability switch values. The random walk value is generated randomly by Python software. While the probability switch value, researchers use 0.8 (Sakti & Putra, 2019). The results are shown in the following table.

Table 6. Random Walk and Probability Switch Values

No.	Random Walk Value	Probability Switch Value
1.	0.37	0.8
2.	0.79	0.8
3.	0.73	0.8
4.	0.60	0.8
5.	0.16	0.8
⋮	⋮	⋮
20.	0.29	0.8

Based on Table 6, the value of random walk < probability switch, which indicates the use of global pollination in FPA. The algorithm performs iterations under the objective function constraints in the search for epsilon and minimum points, and performs maximum iterations when it finds the optimal solution. Using Python on Jupyter Notebook, the iteration result information can be seen as follows.

Table 7. FPA Iteration Results

Iteration	Epsilon	Minimum Points	Objective Function
1	1.90	13	18914.291651
2	1.89	13	17837.757970
3	1.88	13	14101.613183
4	1.87	13	9453.817174
5	1.86	14	6877.361712

⋮	⋮	⋮	⋮
100	1.70	18	5.692261

From the iteration process, it was found that the epsilon value of 1.70 and minimum points 18 at the 100th iteration resulted in the lowest objective function value, which is the optimal parameter of the FPA algorithm.

Cluster Formation Using DBSCAN

Cluster formation was implemented on 438 data consisting of 73 districts/cities and divided into 6 research variables to see the prevalence of leptospirosis in East Java and Central Java Provinces. After the clusters are successfully formed, the next step is to evaluate the clustering results to determine the best number of clusters, using the silhouette coefficient. The best number of clusters can be identified based on the highest silhouette score value. Using Python software on Jupyter Notebook, the results are as follows.

Table 8. Silhouette Coefficient Result

Eps	MinPts	*Silhouette Score*	Number of Clusters (c)
1.70	18	0.1892	2

Based on Table 8, it can be seen that the best silhouette coefficient value is 0.1892 with epsilon 1.70 and minimum points 18, with the optimal number of clusters being 2 clusters. Furthermore, clustering is carried out based on the number of clusters generated previously using Python software on Jupyter Notebook, the results can be seen in the following table.

Table 9. Clustering Results Based on Cluster 0, 1, and Cluster Noise

Clusters	District/City
Cluster 0	Pacitan District, Ponorogo District, Trenggalek District, Tulungagung District, Blitar District, Kediri District, Lumajang District, Banyuwangi District, Probolinggo District, Pasuruan District, Mojokerto District, Jombang District, Nganjuk District, Madiun District, Magetan District, Ngawi District, Bojonegoro District, Tuban District, Lamongan District, Sampang District, Sumenep District, Kebumen District, Magelang District.
Cluster 1	Purbalingga District, Banjarnegara District, Wonosobo District, Boyolali Distr Klaten District, Sukoharjo District, Karanganyar District, Sragen District, Blor District, Rembang District, Kudus District, Jepara District, Demak District, Semarang District, Temanggung District, Kendal District, Batang District, Pekalongan District, Pemalang District, Tegal District.
Cluster Noise	Malang District, Jember District, Bondowoso District, Situbondo District, Sidoarjo District, Gresik District, Bangkalan District, Pamekasan District, Kediri City, Blitar City, Malang City, Probolinggo City, Pasuruan City, Mojokerto City, Madiun City, Surabaya City, Batu City, Cilacap District, Banyumas District, Purworejo District, Wonogiri District, Grobogan District, Pati District, Brebes District, Magelang City, Surakarta City, Salatiga City, Semarang City, Pekalongan City.

Based on Table 9, cluster 0 consists of 23 districts/cities, cluster 1 consists of 20 districts/cities, and 30 districts/cities are detected as noise/outliers. The dominant number of noise/outliers is due to the limitation of core points in reaching these points, which is limited by the epsilon and minimum points parameters. The natural variation in heterogeneous data also affects the

detection of noise/outliers, indicating the clustering algorithm is able to capture the relevant data structure. The following is a mapping of the cluster results using QGIS software.

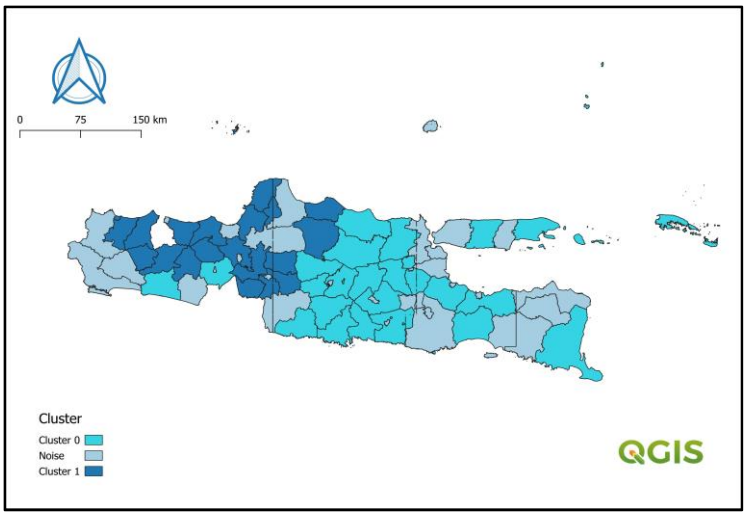

Figure 2. Cluster Mapping Results

Figure 2 visualizes the clustering results for 73 districts/cities, with light blue indicating cluster 0, dark blue for cluster 1, and gray for noise. The characteristics of each cluster are seen based on the average value of the variables, which is shown in the following table. These results were calculated using Python software on Jupyter Notebook.

Table 10. Average Value of Variables in Each Cluster

Variable	Cluster 0	Cluster 1
X_1	-0.556746	0.542259
X_2	-0.523488	-0.353136
X_3	-0.126084	-0.265996
X_4	0.899815	-0.992176
X_5	0.616559	-0.199577
X_6	0.716072	-0.332876

Based on Table 10, cluster 0 has the highest mean values for variables X_3 (Percentage of Houses with Bamboo or Wicker Buildings), X_4 (Percentage of Houses with Distance from Sewers/Waste < 10 metres), X_5 (Number of Farmers), and X_6 (Number of Breeders), indicating the main factors for the spread of leptospirosis in this region are the factors according to these variables. Meanwhile, cluster 1 has the highest mean values for X_1 (Number of Flood Events) and X_2 (Population Density), which are the main factors for the spread of leptospirosis in this cluster area. The Noise cluster contains data that does not belong to any cluster because its characteristics are significantly different from the other data.

The study offers insights for policymakers to develop targeted and efficient health strategies, enabling prioritised surveillance, resource allocation, and intervention in high-risk areas to reduce the leptospirosis burden.

Conclusions and Suggestion

Conclusions: This study aims to identify the pattern of leptospirosis spread in the districts/cities of East Java and Central Java in 2023 through a clustering approach using the DBSCAN method optimised with the Flower Pollination Algorithm (FPA) algorithm. The results showed that the optimal DBSCAN parameters obtained, namely epsilon of 1.70 and

minimum points of 18, produced two clusters and one noise group, with a silhouette coefficient validation value of 0.1892. The clusters formed illustrate different variations in leptospirosis risk characteristics, with 23 districts/cities in cluster 0, 20 districts/cities in cluster 1, and 30 districts/cities identified as noise. This finding is important because it offers a data-driven analytical approach to support more adaptive and spatially-based health policies. The clustering results can be used as a reference in prioritising areas for leptospirosis prevention interventions, planning the distribution of medical resources, and strengthening infectious disease surveillance systems more effectively. By understanding cluster characteristics, policy makers can design more specific treatment strategies according to the risk profile of the region.

Suggestion:

1. Future research should use a larger population and include more complex additional variables, including environmental factors and community behaviour, to improve the accuracy of clustering results.
2. It is also recommended to use a development method of DBSCAN such as Spatial Temporal Density Based Spatial Clustering of Applications with Noise (ST-DBSCAN), which is able to consider spatial and temporal dimensions simultaneously to detect more dynamic patterns of disease spread.
3. Local governments are expected to use the results of this study as a basis for designing preventive and responsive policies, such as the establishment of risk zones, focused allocation of medical personnel, and community education in areas with high-risk cluster characteristics.

Bibliography

Abdussamad, S. N., Astutik, S., & Effendi, A. (2020). Evaluation of Implementation Context Based Clustering In Fuzzy Geographically Weighted Clustering-Particle Swarm Optimization Algorithm. *Jurnal EECCIS*, *14*(1).

Achjar, K. A. H., Agusfina, M., Yesika, R., Aminah, S., Laksono, R. D., Sujati, N. K., … Ifadah, E. (2024). *Penyakit Menular*. PT. Sonpedia Publishing Indonesia.

Afiff, R., Adi, M. S., Saraswati, L. D., & Wuryanto, M. A. (2019). Gambaran Faktor Resiko Leptospirosis Pada Dataran Tinggi Menggunakan Pedoman Kerawanan Leptospirosis Di Dataran Tinggi Dengan Lokasi Penelitian Kabupaten Semarang. *Jurnal Kesehatan Masyarakat*, *7*(3), 2356–3346.

Ali, M., Siswanto, A., & Baehaqi, M. (2024). Flower Polination Algorithm Sebagai Optimalisasi LFC Pada Hybrid Pembangkit Wind-Diesel. *Jurnal FORTECH*, *5*(1), 41–47. https://doi.org/10.56795/fortech.v5i1.5106

Amin, N. F., Garancang, S., & Abunawas, K. (2023). Konsep Umum Populasi Dan Sampel Dalam Penelitian. *JURNAL PILAR: Jurnal Kajiam Islam Kontemporer*, *14*(1).

Ariani, N., & Wahyono, T. Y. M. (2020). Faktor-Faktor yang Mempengaruhi Kejadian Leptospirosis di 2 Kabupaten Lokasi Surveilans Sentinel Leptospirosis Provinsi Banten Tahun 2017-2019. *Jurnal Epidemiologi Kesehatan Indonesia*, *4*.

Aziz, T., & Suwandi, J. F. (2019). Leptospirosis : Intervensi Faktor Resiko Penularan. *Majority*, *8*(1), 232.

Dewi, H. C., & Yudhastuti, R. (2019). Faktor Risiko Kejadian Leptospirosis di Wilayah Kabupaten Gresik (Tahun 2017-2018). *Jurnal Keperawatan Muhammadiyah*, *4*.

Elfaladonna, F., Sartika, D., & Putra, A. M. (2024). Exploratory Data Analysis on the Process of Determining the Relationship between Student Interest and Talent Variables. *SITEKIN: Jurnal Sains, Teknologi Dan Industri*, *21*(2), 418–424.

Harjanto, T. D., Vatresia, A., & Faurina, R. (2021). Analisis Penetapan Skala Prioritas Penanganan Balita Stunting Menggunakan Metode Dbscan Clustering. *Jurnal Rekursif*, *9*(1). Retrieved from http://ejournal.unib.ac.id/index.php/rekursif/30

Husni, S. H., Martini, Suhartono, Budiyono, & Raharjo, M. (2023). Faktor Lingkungan Yang Berpengaruh Terhadap Keberadaan Tikus Serta Identifikasi Bakteri Leptospira sp. di Pemukiman Sekitar Pasar Kota Semarang Tahun 2022. *Jurnal Kesehatan Lingkungan Indonesia*, *22*(2), 134–141. https://doi.org/10.14710/jkli.22.2.134-141

Id, I. D., Astrid, & Mahdiyah, E. (2017). Modifikasi DBSCAN (Density-Based Spatial Clustering With Noise) pada Objek 3 Dimensi. *Jurnal Komputer Terapan*, *3*(1).

Munir, M. M., Wartana, I. M., & Suslistiowati, I. B. (2023). *Integrasi Pembangkit Listrik Tenaga Mikrohidro Pada Sistem Distribusi 20kV Guna Mengurangi Rugi-rugi Daya dan Meningkatkan Profil Tegangan*.

Pribadi, W. W., Yunus, A., & Wiguna, A. S. (2022). Perbandingan Metode K-Means Euclidean Distance Dan Manhattan Distance Pada Penentuan Zonasi Covid-19 Di Kabupaten Malang. *Jurnal Mahasiswa Teknik Informatika*, *6*(2).

Purnama, S. E., & Hartono, B. (2022). Faktor Risiko Kejadian Leptospirosis Di Indonesia: Literature Review. *PREPOTIF Jurnal Kesehatan Masyarakat*, *6*(3).

Puspita, R. N. (2021). Analisis K-Means Cluster Pada Kabupaten/Kota Di Provinsi Banten Berdasarkan Indikator Indeks Pembangunan Manusia. *Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *2*(3).

Putri, M. M., Dewi, C., Siam, E. P., Wijayanti, G. A., Aulia, N., & Nooraeni, R. (2021). Comparison of DBSCAN and K-Means Clustering for Grouping the Village Status in Central Java 2020 Komparasi DBSCAN dan K-Means Clustering pada Pengelompokan Status Desa di Jawa Tengah Tahun 2020. *Jurnal Matematika, Statistika & Komputasi*, *17*(3), 394–404. https://doi.org/10.20956/j.v17i3.11704

Ramadani, A. R. (2021). *Pemodelan Statistical Downscaling menggunakan Regresi Komponen Utama dengan Metode Minimum Vector Variance untuk Pendugaan Curah Hujan. (Studi Kasus: Data Curah Hujan Kabupaten Pangkep)*. Universitas Hasanuddin, Makassar.

Sakti, F. P., & Putra, J. T. (2019). Optimal Reactive Power Dispatch untuk Meminimalkan Rugi Daya Menggunakan Flower Pollination Algorithm. *Jurnal Teknik Elektro*, *11*(2).

Samosir, V. B., Widodo, A. M., Anwar, N., Sekti, B. A., & Erzed, N. (2024). Identifikasi Outlier Menggunakan Teknik Data Mining Clustering Untuk Analisis Data Tracer Study Pada Fakultas Ilmu Komputer Universitas Esa Unggul. *IKRAITH-INFORMATIKA*, *8*(1). https://doi.org/10.37817/ikraith-informatika.v8i1

Sriningsih, M., Hatidja, D., & Prang, J. D. (2018). Penanganan Multikolinearitas Dengan Menggunakan Analisis Regresi Komponen Utama Pada Kasus Impor Beras Di Provinsi Sulut. *Jurnal Ilmiah Sains*, *18*(1).

Warolemba, M. W., Resmawan, & Isa, D. R. (2023). Analisis Cluster Fuzzy C-Means dan Diskriminan untuk Pengelompokan Data Kesejahteraan Rakyat. *Jurnal Sainsmat*, *XII*(2), 141–152. Retrieved from http://ojs.unm.ac.id/index.php/sainsmat

Yang, X.-S. (2012). Flower Pollination Algorithm for Global Optimization. *Unconventional Computation and Natural Computation 2012, Lecture Notes in Computer Science*, *7445*. https://doi.org/10.1007/978-3-642-32894-7_27

Zukhruf, I. A., & Sukendra, D. M. (2020). Analisis Spasial Kasus Leptospirosis Berdasarkan Faktor Epidemologi dan Faktor Risiko Lingkungan. *HIGEIA JOURNAL OF PUBLIC HEALTH RESEARCH AND DEVELOPMENT*, (4).