International Journal of Applied and Advanced Multidisciplinary Research (IJAAMR)
Vol.4, No.5, 2026: 321-336

Density Based Spatial Clustering of Application with Noise on Tuberculosis Cases in Indonesia

Mahrani
Universitas Negeri Makassar
Corresponding Author: Mahrani; mahrani@unm.ac.id

A R T I C L E I N F O

Keywords: DBSCAN Algorithm, Density-Based Clustering, Tuberculosis

Received: 5 March
Revised: 23 April
Accepted: 23 Mei

A B S T R A C T

This study aims to identify the clustering pattern of tuberculosis treatment outcomes in Indonesia in the 2025 research context using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method. This study contributes to public health data mining by applying density-based clustering to classify provincial tuberculosis treatment outcome patterns and detect outlier regions. This quantitative study used secondary tuberculosis data from 34 provinces in Indonesia, observed and analyzed during the 2025 research period. The variables included healed, complete treatment, died, failed treatment, loss to follow-up, and not evaluated indicators. Data were analyzed using normalization, principal component analysis, DBSCAN clustering, and Davies-Bouldin Index validation. The optimal parameters were ε = 0.05 and MinPts = 2, producing two clusters and one noise region. The findings support targeted tuberculosis surveillance and regional health planning

DOI: https://doi.org/10.59890/ijaamr.v4i5.223
ISSN-E: 3026-2437
https://nvlmultitechpublisher.my.id/index.php/ijaamr/index

INTRODUCTION

Clustering is a data mining technique used to identify natural patterns or groups within a dataset based on the similarity of data characteristics (Monshizadeh et al., 2022). In general, clustering methods can be divided into several approaches, including partition-based, hierarchical, grid-based, and density-based clustering. Among these approaches, density-based clustering is particularly useful for identifying groups with irregular shapes, different densities, and the presence of outliers or noise (Martínez-Ratón & Velasco, 2024; Pöelitz & Andrienko, 2010) . One of the most widely used density-based clustering algorithms is Density-Based Spatial Clustering of Applications with Noise, known as DBSCAN. DBSCAN groups data objects based on the density of neighboring points within a certain radius and separates objects that do not belong to any cluster as noise. This characteristic makes DBSCAN relevant for spatial health data because disease distribution often does not follow regular administrative boundaries and may contain extreme or isolated observations.

Tuberculosis remains a serious public health problem in Indonesia. It is an infectious disease caused by Mycobacterium tuberculosis and can be fatal when diagnosis and treatment are delayed (Sanyaolu, 2019). Based on the latest tuberculosis information available during the 2025 research period, Indonesia continues to be one of the countries with the highest tuberculosis burden globally. WHO reported that Indonesia contributed around 10% of global tuberculosis cases in 2024, making it the country with the second-largest share after India. In addition, WHO Indonesia reported that the estimated tuberculosis incidence in Indonesia reached around 1.09 million cases in 2023, while the number of detected and notified cases was approximately 809,000 to 821,000 cases. These figures indicate that tuberculosis control in Indonesia still faces major challenges, especially in case detection, reporting, treatment access, and regional surveillance.

The distribution of tuberculosis cases in Indonesia is not uniform across regions because it may be influenced by population density, mobility, socioeconomic conditions, environmental quality, and access to health services. This uneven spatial distribution creates a practical need for clustering analysis so that areas with high case density and areas considered as noise can be identified more clearly. Previous studies on tuberculosis clustering have commonly used partition-based methods such as K-Means and Fuzzy K-Means (Rochman et al., 2022a). However, these methods generally require the number of clusters to be determined in advance and may be less suitable for spatial data containing irregular cluster shapes and outliers. DBSCAN offers a different analytical advantage because it does not require the initial number of clusters and can detect noise in the dataset. Therefore, the contribution of this paper lies in applying the DBSCAN method to tuberculosis cases in Indonesia as a spatial data mining approach to enrich public health analysis, particularly in identifying dense case areas and separating possible outlier regions. This study aims to analyze the spatial clustering pattern of tuberculosis cases in Indonesia using the DBSCAN method and to describe how the resulting clusters can support more targeted tuberculosis surveillance and control strategies.

LITERATURE REVIEW

Density-Based Clustering Theory

Density-based clustering is a clustering approach that groups data objects based on the density level of data points in a particular area. This method identifies clusters by observing how closely data points are located to one another. Unlike partition-based clustering, density-based clustering does not require the number of clusters to be determined before the analysis. This approach is useful for data that have irregular cluster shapes, different density levels, and the presence of outliers or noise. Therefore, density-based clustering is suitable for spatial data analysis because spatial phenomena often do not form regular or balanced patterns across regions(T. fan Zhang et al., 2022).

One of the most widely used density-based clustering algorithms is Density-Based Spatial Clustering of Applications with Noise, known as DBSCAN. DBSCAN forms clusters using two main parameters, namely epsilon and minimum points (Zhu et al., 2021). Epsilon refers to the maximum radius or distance used to determine neighboring points, while minimum points refer to the minimum number of data points required to form a dense area. Data points located in dense areas are grouped into clusters, while data points that do not meet the density requirement are categorized as noise. This characteristic makes DBSCAN relevant for identifying spatial patterns of tuberculosis cases because disease cases may be concentrated in certain areas and scattered irregularly in others (Monshizadeh et al., 2022).

Previous studies have shown that clustering methods can be applied to tuberculosis data. Wardani et al. (2019) used the K-Means method to cluster tuberculosis cases in children, while Rochman, Miswanto, and Suprajitno (2022) compared Fuzzy K-Means and K-Means methods for tuberculosis disease clustering. These studies indicate that clustering can help describe the distribution pattern of tuberculosis cases. However, K-Means and Fuzzy K-Means generally require the number of clusters to be determined at the beginning and are less effective when the data contain irregular shapes or outliers. Based on this limitation, DBSCAN is considered more appropriate because it can identify clusters based on density and separate noise from the main cluster structure.

H1: The DBSCAN method can identify density-based clusters of tuberculosis cases in Indonesia.

Spatial Epidemiology Theory

Spatial epidemiology is a theory and analytical approach used to study the geographical distribution of disease and its relationship with population, environmental, and health service factors. This theory assumes that disease cases are not randomly distributed across space, but may form certain patterns due to differences in population density, mobility, socioeconomic conditions, environmental quality, and access to health services. In public health studies, spatial epidemiology is important because it helps identify areas with a higher disease burden and supports more targeted health intervention (R. Zhang et al., 2022).

In the context of tuberculosis, spatial epidemiology is relevant because tuberculosis transmission and case detection may differ between regions. Areas with high population density, poor environmental conditions, limited access to health services, or high mobility may have different tuberculosis case patterns compared to areas with better health infrastructure (R. Zhang et al., 2022). Therefore, tuberculosis should not only be analyzed through total case numbers, but also through spatial distribution patterns. By using a spatial clustering approach, regions with similar tuberculosis case density can be grouped and interpreted more clearly.

DBSCAN supports the application of spatial epidemiology because it can detect dense case areas and identify regions that do not belong to any cluster. In tuberculosis surveillance, cluster areas may indicate regions requiring greater attention, while noise areas may reflect isolated cases or unusual distribution patterns. Thus, the use of DBSCAN can enrich spatial epidemiological analysis by providing information about the density and irregularity of tuberculosis distribution in Indonesia (Daszykowski & Walczak, 2009).

H2: Tuberculosis cases in Indonesia form spatial distribution patterns that can be classified into clusters and noise using the DBSCAN method.

Tuberculosis Surveillance and Regional Health Planning

Tuberculosis surveillance is an important component of disease control because it provides information about case distribution, case detection, treatment outcomes, and intervention priorities. Effective surveillance requires not only accurate case recording but also proper data analysis to identify regions with high disease concentration. In Indonesia, tuberculosis remains a major public health problem, and its distribution is not uniform across provinces or regions. This condition creates a need for analytical methods that can support regional mapping and decision-making (Rochman et al., 2022b).

Clustering analysis can help transform tuberculosis data into useful information for public health planning. Through clustering, regions with similar case characteristics can be grouped, making it easier to identify priority areas for intervention. For example, areas included in high-density clusters may require stronger case finding, contact tracing, treatment monitoring, and health education programs. Meanwhile, regions identified as noise may require further investigation because they may represent isolated cases, unusual reporting patterns, or areas with specific local characteristics (R. Zhang et al., 2022).

The use of DBSCAN in tuberculosis surveillance provides an analytical contribution because it does not only classify regions into groups but also identifies noise. This is important because public health data often contain irregular patterns and extreme observations. Therefore, DBSCAN can support more targeted tuberculosis surveillance and regional health planning by identifying areas with high case density and distinguishing them from areas with irregular patterns.

H3: The clustering results produced by DBSCAN can support tuberculosis surveillance and regional health planning in Indonesia.

Conceptual Framework

The conceptual framework of this study explains the relationship between tuberculosis case data, spatial characteristics, DBSCAN parameters, clustering results, and public health interpretation (Martínez-Ratón & Velasco, 2024). This study begins with tuberculosis case data in Indonesia, which are analyzed based on spatial and case-density characteristics. The DBSCAN method is then applied using epsilon and minimum points as the main parameters (Loh & Park, 2014). Through the density-based clustering process, the data are classified into cluster areas and noise areas. The results are then interpreted to identify tuberculosis distribution patterns and to support more targeted tuberculosis surveillance and control strategies.

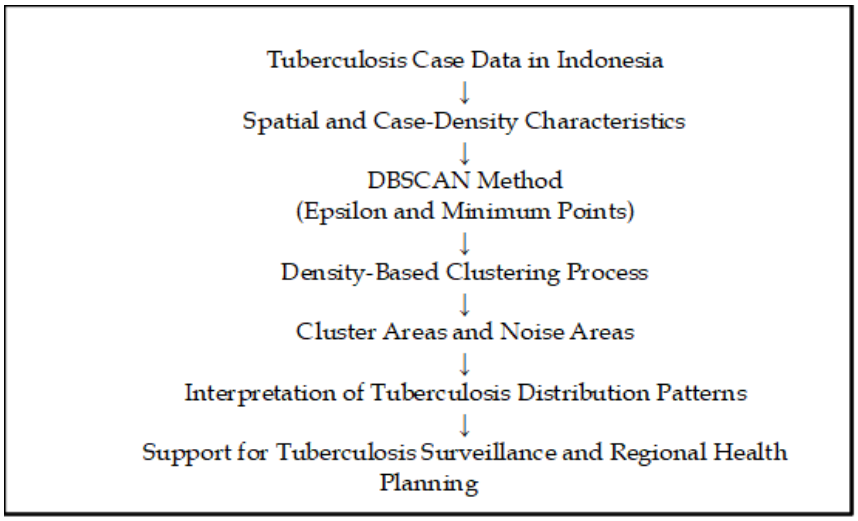

Figure 1. Conceptual Framework

METHODOLOGY

This study used a quantitative data mining approach to analyze tuberculosis cases in Indonesia using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method (Latifi-Pakdehi & Daneshpour, 2021). The population of this study consisted of tuberculosis case data in Indonesia, while the sample included provincial-level tuberculosis data obtained from official sources available during the 2025 research period. Each province was treated as one unit of analysis. The variables used in this study consisted of tuberculosis-related indicators and regional attributes relevant to clustering analysis. Before clustering, the data were preprocessed through normalization, multicollinearity detection, and dimensional reduction . Data normalization was conducted because cluster analysis is sensitive to differences in variable scales. Variables with large value differences may dominate distance calculation and produce biased clusters. Therefore, the data were standardized using the Z-score transformation as follows:

$$Z_{pa} = \frac{x_{pa} - \bar{x}_a}{S_a}, p = 1,2, \dots, n \text{ dan } a = 1,2 \dots, l \quad (1)$$

where $Z_{pa}$is the standardized value of observation $p$on variable a, $x_{pa}$is the original value, $\bar{x}_a$is the mean of variable a, and $S_a$is the standard deviation of variable (Dahmouni et al., 2018).

Multicollinearity detection was carried out by examining the correlation coefficient between variables. This step was used to identify variables with strong relationships that may contain overlapping information. The correlation coefficient was calculated as follows: (Anderson et al., 2014)

$$r_{x_a,x_b} = \frac{n\left(\sum_{p=1}^{n} x_{pa}x_{pb}\right)-\left(\sum_{p=1}^{n} x_{pa}\right).\left(\sum_{p=1}^{n} x_{pb}\right)}{\sqrt{n\left(\sum_{p=1}^{n} {x_{pa}}^2\right)-\left(\sum_{p=1}^{n} x_{pa}\right)^2}.\sqrt{n\left(\sum_{p=1}^{n} {x_{pb}}^2\right)-\left(\sum_{p=1}^{n} x_{pb}\right)^2}}, \text{ p= 1,2,…,} \quad (2)$$

Principal Component Analysis (PCA) was then applied to reduce data dimensions and generate new variables that represent the main information contained in the original variables. PCA was used when the variables showed multicollinearity (Spiegel & Stephens, 2008). The adequacy of PCA was evaluated using Bartlett's test and the Kaiser-Meyer-Olkin (KMO) test. Bartlett's test was used to determine whether the variables were sufficiently correlated, while the KMO test was used to measure sampling adequacy. A KMO value greater than 0.5 indicates that factor analysis is appropriate. The factor score was calculated using the following equation:

$$F_i = w_{i1}X_1 + w_{i2}X_2 + \cdots + w_{ij}X_j \quad (3)$$

where $F_i$is the factor score, $w_{ij}$is the factor weight, and $X_j$is the standardized variable.

After preprocessing, clustering analysis was performed using the DBSCAN algorithm. DBSCAN was selected because it can identify clusters based on local density and detect noise or outlier observations (R. Zhang et al., 2022). The algorithm uses two main parameters, namely epsilon $(\varepsilon)$and minimum points $(MinPts)$. Epsilon determines the maximum distance between neighboring points, while $MinPts$determines the minimum number of observations required to form a cluster. The distance between observations was calculated using Euclidean distance. Next, we calculate the distance for all observational data using (T. fan Zhang et al., 2022)

$$d(p,q) = \left\|x_p - x_q\right\|_2 = \sqrt{\sum_{a=1}^{l}\left(x_{pa} - x_{qa}\right)^2}, \text{ p,q= 1,2,3,…,n} \quad (4)$$

where $d(p,q)$is the distance between observations $p$and q, and $x_{pa}$and $x_{qa}$are the values of variable $a$for observations $p$and q.

The clustering results were validated using the Davies-Bouldin Index (DBI). DBI was used to evaluate cluster quality based on cluster compactness and separation. A smaller DBI value indicates better clustering performance because members within the same cluster are more homogeneous and clusters are more clearly separated from one another. The DBI formula is expressed as follows: (Gnimassoun et al., 2024)

$$DBI = \frac{1}{n}\sum_{i=1}^{n} \max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_i)}\right) \quad (5)$$

where $k$is the number of clusters, $\sigma_i$and $\sigma_j$are the average distances of objects to the centroid of clusters $i$and j, and $d(c_i, c_j)$is the distance between cluster centroids. The data analysis was conducted using R/Python to perform normalization, PCA, DBSCAN clustering, visualization, and cluster validation.

RESULT

Data Normalization

Data normalization was conducted before clustering analysis to avoid bias caused by differences in variable scales. Since DBSCAN uses distance-based calculation, variables with larger numerical ranges may dominate the clustering process if the data are not standardized. In this study, normalization was carried out using the Z-score transformation as stated in Equation (1). The normalized data were then used as the input for the next analysis stage.

Table 1. Data Normalization

No	Z_1	Z_2	Z_3	⋯	Z_6
1	0.1783995	1.00000	0.0515873	⋯	0.006481481
2	0.2856878	0.9087302	0.0515873	⋯	0.001058201
3	0.4033730	0.7883598	0.06084656	⋯	0.008597884
⋮	⋮	⋮	⋮	⋱	⋮
34	0.7118915	0.7116402	0.07671958	⋯	0.017857143

Based on Table 1, all variables were transformed into standardized values so that each variable had a comparable scale. This process made the data more suitable for distance-based clustering and reduced the possibility of biased cluster formation.

Multicollinearity Test

The multicollinearity test was conducted to determine the correlation between variables used in the clustering process. This stage is important because highly correlated variables may contain overlapping information and may affect the clustering structure. The correlation matrix between variables is presented in Table 2.

Table 2. Matriks Korelasi

	Healed	Complete Treatment	Died	Failed	Loss to followup	Not evaluated
Healed	1	-0.4489	0.2781	-0.0303	0.1427	0.3057
Complete Treatment	-0.4489	1	0.3798	-0.2193	-0.1580	-0.1105
Died	0.2781	-0.3798	1	0.3514	0.0781	-0.2789
Failed	-0.0303	-0.3798	0.3514	1	0.4888	-0.0028
Loss to followup	0.1427	-0.2193	0.0781	0.4888	1	0.1038
Not evaluated	0.3057	-0.1580	-0.2789	-0.0028	0.1038	1

Based on Table 2, the highest correlation coefficient was found between Failed and Loss to Follow-up, with a value of 0.4888. This value indicates a moderate relationship, not a very strong correlation. Therefore, the variables did

not show severe multicollinearity. However, dimensional reduction through Principal Component Analysis was still conducted to simplify the data structure and obtain more representative components for clustering.

Principal Component Analysis

Principal Component Analysis was applied to reduce the dimensions of the dataset and to retain the main information from the original variables. Before PCA was performed, Bartlett's test and the Kaiser-Meyer-Olkin test were used to examine the suitability of the data for factor analysis. The KMO value obtained was 0.5675088, which is greater than 0.5. This result indicates that the data were adequate for factor analysis. The component values are presented in Table

Table 3. Component Value

Faktor	Eigenvalues	Proportion of Variance	Cumulative Proportion
1	2.03	0.7730	0.7730
2	1.38	0,2301	0,5682
3	1.21	0,2011	0,7693
4	0.58	0,09634	0,86560
5	0.52	0,08607	0,95167
6	0.29	0,04833	1.00

Based on Table 3, the first component had an eigenvalue of 2.03 and explained 77.30% of the variance. Since this value exceeded the minimum cumulative variance criterion of 60%, the first principal component was considered sufficient to represent the main information in the dataset. Therefore, the DBSCAN clustering process was conducted using the principal component score as the clustering input.

DBSCAN Clustering Result

The DBSCAN algorithm was applied to identify density-based clusters of tuberculosis patient status in Indonesia. This method was selected because it can detect clusters with irregular patterns and identify outlier or noise regions. The main parameters used in DBSCAN were epsilon (ε)and minimum points (MinPts). Based on the K-nearest neighbor distance plot, the epsilon value was set at 0.05, while the MinPtsvalue was set at 2. The clustering visualization is presented in Figure 1 and Figure 2.

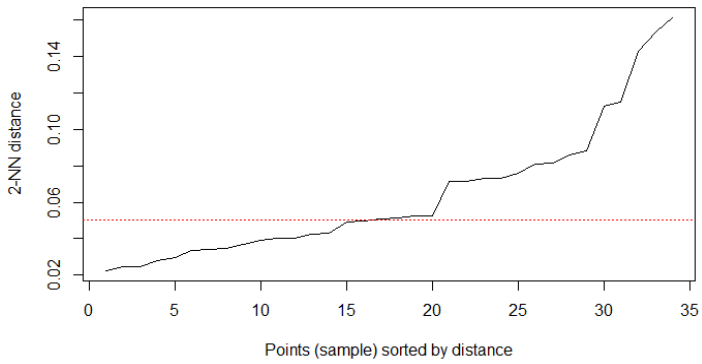

Figure 2. Grafik KNN-displot

Figure 1 presents the K-nearest neighbor distance plot used to determine the epsilon (ε)value in the DBSCAN clustering algorithm. The KNN-distance plot helps identify the most appropriate neighborhood radius by observing the point where the distance curve begins to show a significant change or elbow pattern. Based on the graph, the epsilon value was determined at 0.05. This value indicates the maximum distance between observations that can still be considered neighbors in the clustering process. After the epsilon value was obtained, the DBSCAN analysis was continued using MinPts = 2, meaning that at least two observations within the radius of 0.05 were required to form a dense region. The combination of ε = 0.05 and MinPts = 2 was then used to classify the provinces into cluster groups and noise regions based on the similarity of tuberculosis treatment outcome characteristics.

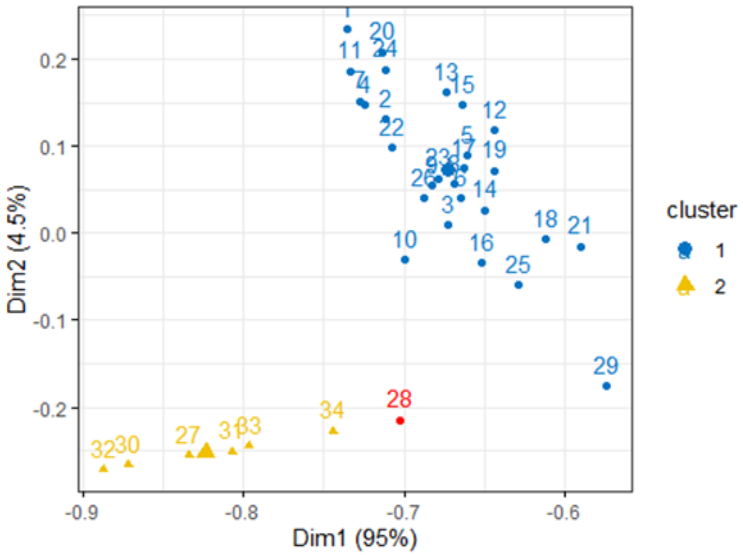

Figure 3. Cluster Plot

The visualization in Figure.2 displays the clusters obtained by applying the DBSCAN algorithm with an Epsilon value of 0.05 and a MinPts value of 2. The cluster analysis shows two distinct clusters. Cluster 1 consists of 27 elements, while cluster 6 contains only one element. In addition, there is one element that does not belong to cluster 1 or cluster 2, indicating that the element is considered noise.

Table 4. Cluster Result

Cluster	Sample
0	Bali
1	Aceh, North Sumatra, West Sumatra, Riau, Riau Islands, Jambi, South Sumatra, Bangka Belitung Islands, Bengkulu, Lampung, Banten, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Gorontalo, Central Sulawesi, South Sulawesi, West Sulawesi, West Nusa Tenggara

2	Southeast Sulawesi, East Nusa Tenggara, Maluku, North Maluku, Papua, West Papua

Based on Table 4, shows that there is one region included in the outlier category, namely Bali with a cure rate of 0.5154%, there are 27 regions included in the cluster 1 category, namely Aceh, North Sumatra, West Sumatra, Riau, Riau Islands, Jambi, South Sumatra, Bangka Belitung Islands, Bengkulu, Lampung, Banten, DKI Jakarta, West Java, Central Java, DI, East Java, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Gorontalo, Central Sulawesi, South Sulawesi, West Sulawesi, West Nusa Tenggara with a cure rate of 0.241% and there are 6 regions included in cluster 2 namely Southeast Sulawesi, East Nusa Tenggara, Maluku, North Maluku, Papua, West Papua. Cluster 0 is close to cluster 2, but far from cluster 1 with a cure rate of 0.604%.

Validasi Cluster

The cluster validation method used is the Davies-Bouldin Index (DBI), where optimal clustering is that which produces a DBI value close to zero (non-negative, greater than or equal to zero) [3]. After clustering with the DBSCAN technique, cluster validation is performed by checking the Davies-Bouldin Index (DBI) value for each parameter combination.

Table 6. DBI Value

	MinPts						
ε	2	3	4	5	6	7	8
0.01	1.211	1.372	1.702.	1.892	1.567	NA	NA
0.02	0.922	1.345	NA	NA	NA	NA	NA
0.03	1.321	1.109	NA	NA	NA	NA	NA
0.04	1.198	1.421	NA	NA	NA	NA	NA
0.05	0.450	1.311	1.034	1.598	0.854	0.988	0.871
0.06	0.830	0.722	0.982	0.789	0.901	NA	NA
0.07	NA	NA	NA	NA	NA	NA	NA
0.08	NA	NA	NA	NA	NA	NA	NA

Based on Table 6, the best parameter combination was obtained at ε = 0.05 and MinPts = 2, with the smallest DBI value of 0.450. This result indicates that this parameter combination produced the most optimal cluster structure compared with the other tested combinations. The NA values indicate that the DBI could not be calculated because some parameter combinations did not produce valid cluster structures, such as clusters with no members or zero variance. Therefore, the DBSCAN model with ε = 0.05 and MinPts = 2 was selected as the final clustering model. The result shows that tuberculosis patient-status data in Indonesia can be grouped into two main clusters and one noise region. This finding demonstrates that DBSCAN is able to identify density-based patterns and detect outlier regions in tuberculosis-related data.

DISCUSSION

The findings of this study show that the DBSCAN method is able to identify density-based patterns in tuberculosis patient-status data in Indonesia. The clustering process produced two main clusters and one noise region. This result indicates that tuberculosis patient-status characteristics across Indonesian provinces are not fully homogeneous. Some provinces share similar patterns and are grouped into the same cluster, while one province shows a different pattern and is identified as noise. This finding supports the basic assumption of density-based clustering, which states that data objects located in dense areas can form clusters, while objects that do not meet the density requirement are separated as noise.

The selection of ε = 0.05 and MinPts = 2 as the best DBSCAN parameters indicates that the clustering structure was formed using a relatively small neighborhood radius and a minimum of two observations to create a dense region. This parameter combination produced the lowest Davies-Bouldin Index value, namely 0.450, which means that the resulting clusters had better compactness and separation compared with other parameter combinations tested in the study. A lower DBI value reflects a better clustering structure because members within the same cluster tend to be more similar, while members from different clusters are more clearly separated. Therefore, the chosen DBSCAN model can be considered appropriate for identifying tuberculosis patient-status patterns in this dataset.

The result showing Cluster 1 as the largest cluster with 27 provinces suggests that most provinces in Indonesia have relatively similar tuberculosis patient-status characteristics. This similarity may indicate that many provinces share comparable treatment outcome patterns, such as healed cases, complete treatment, death, treatment failure, loss to follow-up, and unevaluated cases. From a public health perspective, this finding implies that a large number of provinces may face similar tuberculosis management conditions. Therefore, national tuberculosis control strategies may still be relevant for provinces in this cluster, although local adjustments remain necessary.

Meanwhile, Cluster 2 consists of six provinces, namely Southeast Sulawesi, East Nusa Tenggara, Maluku, North Maluku, Papua, and West Papua. The separation of these provinces from the larger cluster indicates that they have different tuberculosis patient-status characteristics compared with most other provinces. This finding is important because these regions generally have geographical challenges, dispersed populations, and unequal access to health services. In the context of tuberculosis control, provinces in this cluster may require more specific surveillance and intervention strategies. The clustering result suggests that public health policies should not treat all regions in the same way, because some provinces may have distinct epidemiological and service-access patterns.

The identification of Bali as a noise region is also an important finding. In DBSCAN, noise does not simply mean an error in the data; rather, it indicates an observation that does not have sufficient density similarity with other observations. This means that Bali has tuberculosis patient-status characteristics

that differ from the dominant patterns found in other provinces. This result requires further investigation because a noise region may reflect unique local conditions, differences in reporting quality, health service performance, treatment outcomes, population mobility, or other contextual factors. Therefore, the appearance of Bali as noise should be interpreted as an analytical signal that needs further public health explanation, not as a meaningless outlier.

These findings also show the advantage of DBSCAN compared with partition-based clustering methods such as K-Means. Previous tuberculosis clustering studies often used K-Means or Fuzzy K-Means, which require the number of clusters to be determined before analysis. However, tuberculosis-related data may contain irregular patterns and isolated observations. In this situation, DBSCAN is more flexible because it can form clusters based on density and identify noise without requiring the researcher to determine the number of clusters at the beginning. This strengthens the methodological contribution of the present study, especially in the application of density-based clustering for tuberculosis surveillance in Indonesia.

From an epidemiological perspective, the results confirm that tuberculosis data should not only be analyzed based on total cases, but also through patterns of similarity among regions. Clustering helps transform statistical data into more useful information for decision-making. Provinces within the same cluster can be interpreted as having similar patient-status profiles and may be considered for similar intervention strategies. On the other hand, provinces classified into different clusters or noise areas require more specific attention because they may represent different levels of treatment success, case management problems, or health service challenges.

The findings have practical implications for tuberculosis surveillance and regional health planning. The clustering results can help policymakers identify groups of provinces that need similar intervention models and distinguish provinces that require special attention. For provinces in the largest cluster, intervention may focus on strengthening existing tuberculosis control programs at a broad scale. For provinces in the smaller cluster, strategies may need to consider regional barriers such as geography, access to health facilities, and continuity of treatment. For noise regions, further investigation is needed to understand why their tuberculosis patient-status pattern differs from other provinces.

Overall, the results indicate that DBSCAN is useful for identifying density-based tuberculosis patterns in Indonesia. The method does not only classify provinces into clusters but also detects regions with unusual characteristics. This supports the contribution of the study to spatial and public health data mining by showing that DBSCAN can be used as an analytical tool for tuberculosis surveillance. The findings imply that tuberculosis control strategies should be supported by data-driven regional classification so that interventions can be more targeted, efficient, and responsive to local characteristics.

CONCLUSIONS AND RECOMMENDATIONS

This 2025 study analyzed provincial tuberculosis treatment outcome data in Indonesia to identify density-based clustering patterns using the DBSCAN method. The descriptive results show that the average proportion of cured patients was 31.86%, complete treatment was 61.85%, death was 4.84%, treatment failure was 0.27%, loss to follow-up was 6.27%, and not evaluated cases were 2.05%. Among the evaluated provinces, East Java, Central Java, and West Java showed the highest tuberculosis treatment outcome values based on the six observed indicators.

The DBSCAN analysis conducted in this study showed that the optimal parameter combination was $\varepsilon = 0.05$ and MinPts = 2, with the lowest Davies-Bouldin Index value of 0.450. This result indicates that the selected parameter combination produced the most reliable cluster structure among the tested alternatives. The analysis generated two main clusters and one noise region. Bali was identified as the noise region, Cluster 1 consisted of 27 provinces with moderate treatment outcome characteristics, and Cluster 2 consisted of 6 provinces with relatively high treatment outcome characteristics. These findings demonstrate that DBSCAN can effectively classify Indonesian provinces based on tuberculosis treatment outcome similarities and detect regions with distinctive patterns that require further public health attention.

Recommendations

The results of this study can be implemented as a supporting tool for tuberculosis surveillance and regional health planning in Indonesia. Provinces within the same cluster may be given similar intervention strategies because they share comparable treatment outcome characteristics. Provinces in Cluster 1 require continued strengthening of general tuberculosis control programs, especially treatment completion, follow-up monitoring, and evaluation of treatment outcomes. Provinces in Cluster 2 should be studied further to identify factors that contribute to better treatment outcomes, so that good practices from these regions can be adapted to other provinces.

The identification of Bali as a noise/outlier region indicates the need for further investigation. This result may reflect a different treatment outcome pattern, reporting characteristic, health service condition, or local epidemiological factor. Therefore, policymakers should not only focus on the largest cluster but also pay attention to outlier regions because they may contain important information for improving tuberculosis control strategies. Future research is recommended to use more detailed data at the district or city level and to include additional variables such as population density, health facility access, poverty level, environmental conditions, and treatment success rate to obtain more precise clustering results.

FUTURE STUDY

This study has several limitations that need to be considered. First, the analysis was conducted using provincial-level tuberculosis data, so the clustering results only describe general patterns at the province level. This may limit the ability to identify more specific tuberculosis distribution patterns at the district

or city level. Second, the variables used in this study were limited to tuberculosis treatment outcome indicators, including healed, complete treatment, died, failed treatment, loss to follow-up, and not evaluated cases. Other important factors such as population density, poverty level, environmental conditions, health facility availability, and access to tuberculosis services were not included in the clustering process.

Another limitation is related to the DBSCAN parameter selection. The clustering result was strongly influenced by the choice of epsilon and MinPts. In this study, the best parameter combination was $\varepsilon = 0.05$ and MinPts = 2, which produced the lowest Davies-Bouldin Index value of 0.450. However, different datasets, variables, or spatial scales may produce different optimal parameter values. Therefore, the result should be interpreted according to the characteristics of the dataset used in this study.

Future research is recommended to use more detailed tuberculosis data at the district or city level to obtain more precise clustering results. Further studies may also include additional epidemiological, demographic, socioeconomic, and health service variables to provide a more comprehensive explanation of tuberculosis clustering patterns. In addition, future researchers can compare DBSCAN with other clustering methods, such as K-Means, Fuzzy C-Means, hierarchical clustering, or other density-based algorithms, to evaluate which method produces the most accurate and interpretable clustering results for tuberculosis surveillance in Indonesia.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to all parties who provided suggestions, guidance, and constructive input during the preparation of this paper. Appreciation is also given to the institution, lecturers, and colleagues who supported the completion of this research. The authors acknowledge the use of tuberculosis data from official health sources as the basis for analysis. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., & Cochran, J. J. (2014). Statistics for business and economics. www.nelson.com

Dahmouni, A., El Moutaouakil, K., & Satori, K. (2018). Clustering and jarque-bera normality test to face recognition. Procedia Computer Science, 127, 246–255. https://doi.org/10.1016/j.procs.2018.01.120

Daszykowski, M., & Walczak, B. (2009). Density-Based Clustering Methods.

Gnimassoun, J. E., Ricky N'DRI, A. K., & Legrand KOFFI, Dagou Dangui Augustin Sylvain. (2024). Efficient Workflow Scheduling for Minimizing Data Transfers and Enhancing Resource Utilization in Cloud IaaS Platforms. INTERNATIONAL JOURNAL OF

MATHEMATICS AND COMPUTER RESEARCH, 12(11). https://doi.org/10.47191/ijmcr/v12i11.01

Latifi-Pakdehi, A., & Daneshpour, N. (2021). DBHC: A DBSCAN-based hierarchical clustering algorithm. Data and Knowledge Engineering, 135. https://doi.org/10.1016/j.datak.2021.101922

Loh, W. K., & Park, Y. H. (2014). A survey on density-based clustering algorithms. Lecture Notes in Electrical Engineering, 280 LNEE, 775–780. https://doi.org/10.1007/978-3-642-41671-2_98

Martínez-Ratón, Y., & Velasco, E. (2024). Density-functional theory for clustering of two-dimensional hard particle fluids. Journal of Molecular Liquids, 397. https://doi.org/10.1016/j.molliq.2024.124044

Monshizadeh, M., Khatri, V., Kantola, R., & Yan, Z. (2022). A deep density based and self-determining clustering approach to label unknown traffic. Journal of Network and Computer Applications, 207. https://doi.org/10.1016/j.jnca.2022.103513

Pöelitz, C., & Andrienko, N. (2010). Finding arbitrary shaped clusters with related extents in space and time. www.flickr.com

Rochman, E. M. S., Miswanto, & Suprajitno, H. (2022a). COMPARISON OF CLUSTERING IN TUBERCULOSIS USING FUZZY C-MEANS AND K-MEANS METHODS. Communications in Mathematical Biology and Neuroscience, 2022. https://doi.org/10.28919/cmbn/7335

Rochman, E. M. S., Miswanto, & Suprajitno, H. (2022b). COMPARISON OF CLUSTERING IN TUBERCULOSIS USING FUZZY C-MEANS AND K-MEANS METHODS. Communications in Mathematical Biology and Neuroscience, 2022. https://doi.org/10.28919/cmbn/7335

Sanyaolu, A. (2019). Tuberculosis: A Review of Current Trends. Epidemiology International Journal, 3(2). https://doi.org/10.23880/eij-16000123

Spiegel, M. R., & Stephens, L. J. (2008). Schaum's outline of statistics (4th ed.). https://doi.org/10.1036/0071485848

Zhang, R., Qiu, J., Guo, M., Cui, H., & Chen, X. (2022). An Adjusting Strategy after DBSCAN. IFAC-PapersOnLine, 55(3), 219–222. https://doi.org/10.1016/j.ifacol.2022.05.038

Zhang, T. fan, Li, Z., Yuan, Q., & Wang, Y. ning. (2022). A spatial distance-based spatial clustering algorithm for sparse image data. Alexandria Engineering Journal, 61(12), 12609–12622. https://doi.org/10.1016/j.aej.2022.06.045

Zhu, Q., Tang, X., & Elahi, A. (2021). Application of the novel harmony search optimization algorithm for DBSCAN clustering. Expert Systems with Applications, 178. https://doi.org/10.1016/j.eswa.2021.115054