



Comparison of K-Means and K-Medoids Clustering Algorithms for Mapping the Spread of Tuberculosis Disease in Central Java Province

Fitria Mahmudah*, Putriaji Hendikawati

Department of Mathematics, Faculty of Mathematics and Natural Science,
Universitas Negeri Semarang, Semarang, 50229, Indonesia

Article Info

Article History:

Received 24-06-2025

Accepted 20-04-2026

Published 30-04-2026

Keywords:

*K-Means, K-Medoids,
Tuberculosis, Davies-
Bouldin Index*

Abstract

Tuberculosis (TB) is a major public health problem: Indonesia accounts for about 10% of the estimated TB cases in the world, second only to India, and Central Java is among the three provinces with the highest burden in the country, so districts with the greatest need must be identified for targeted control. This study maps the spread of TB in Central Java by clustering 35 districts/cities on ten indicators of TB cases and outcomes in 2024, standardized by z-scores. K-Means was compared with K-Medoids, which is more robust to outliers, using Euclidean, Manhattan, and Minkowski distances; $k = 3$ was chosen from the elbow and silhouette methods; and the Davies-Bouldin Index (DBI) was used for validation. K-Means with Manhattan distance had the lowest DBI (0.7789), against 0.7822 for K-Means with Euclidean and Minkowski distances and 0.7803 to 0.7950 for K-Medoids, so the differences among the methods are small. The clusters comprise 22 low, 9 moderate, and 4 high districts/cities; the high cluster (Banyumas, Brebes, Tegal Regency, and Semarang City) has z-scores above 1.9 in eight of the ten indicators. Priority should be given to the high cluster for service capacity, treatment access, and case finding, and a watch list should be kept for the moderate cluster; future analyses should use incidence rates.

How to cite:

Mahmudah, F., & Hendikawati, P. (2026). Comparison of K-means and K-medoids clustering algorithms for mapping the spread of tuberculosis disease in Central Java Province. *UNNES Journal of Mathematics*. 15(1): 14-24

1. Introduction

In today's digital era, advances in information technology have driven the availability and growth of large amounts of data (big data) from various fields of life. This massive data not only reflects information in large volumes, but also has complex characteristics such as speed, variety, reliability, and high value. Big data is a collection of data obtained from various sources, to obtain information from the data, various analyses are carried out with data mining (Syamsu et al., 2021). Big data has five main characteristics known as 5V: volume (data size), velocity (data speed), variety (data variation), veracity (reliability and authenticity of data), and value (benefit from the data set).

One of the widely used techniques in data mining is Clustering or data grouping. Clustering aims to group objects based on attribute similarities so that data in one group has similar characteristics. This technique is widely used in various fields, including economics, health, and demography (Dani et al., 2019). There are two approaches in cluster analysis, namely hierarchical and non-hierarchical methods. The hierarchical method is used to form a hierarchical structure of a data set based on its characteristics. Meanwhile, the non-hierarchical method is an approach that groups data into a number of predetermined clusters. The two most commonly used non-hierarchical methods are K-Means and K-Medoids. K-Means uses centroids as cluster centers, while K-Medoids uses medoids that are more resistant to outlier data (Schubert & Rousseeuw, 2019).

In the health sector, clustering is relevant for studying the spread of infectious diseases such as tuberculosis (TB). TB is transmitted mainly through the air, for example by coughing, and it can also attack other organs, known as extrapulmonary tuberculosis (Alam et al., 2019). TB remains a serious public health problem both globally and nationally. Indonesia accounts for about 10% of the estimated TB cases in the world, second only to India, and more than 800,000 cases were reported in 2023 (WHO, 2024). In Central Java Province, the number of TB cases continues to increase from year to year, making it one of the three provinces with the highest burden in Indonesia (Ministry of Health of the Republic of Indonesia, 2023).

The concrete problem addressed here is that TB control resources of the 35 districts/cities of Central Java must be allocated, but the districts differ in the number of cases, in the age and sex of patients, and in treatment outcomes and drug-resistant (RO) TB, so that a single indicator does not show which districts need the most attention. The urgency is that decision makers need a grouping of districts that uses all these indicators together, and a method that has been compared and validated, to decide where to strengthen service capacity, treatment access, and case finding first.

Clustering was used because it groups districts by the similarity of many indicators without labels, which suits this unsupervised problem (Priyatman et al., 2019; Loklomin et al., 2022). K-Means and K-Medoids were compared because they are the most commonly used partitioning methods, and because they differ in the cluster center: K-Means uses the mean, which is sensitive to extreme values, whereas K-Medoids uses an actual observation, which is more robust to outliers (Schubert & Rousseeuw, 2019). This matters for TB data in which a few districts have very high numbers of cases. Three distances were compared because the distance defines similarity: Euclidean is the standard choice, Manhattan is less dominated by a single large difference, and Minkowski generalizes both. The Davies-Bouldin Index was used because it evaluates cohesion and separation without reference labels, and the elbow and silhouette methods were used to choose the number of clusters.

Previous studies compared the two algorithms or the distances on other data. Utomo (2021) found that K-Means had a lower DBI than K-Medoids for the spread of Covid-19 in Indonesia, and Febrian and Novianti (2024) found Manhattan distance better than Euclidean distance for K-Means on heart disease data. For TB, Loklomin et al. (2022) clustered districts of North Maluku by TB determinants, and Faidah et al. (2025) compared K-Means and DBSCAN for TB mapping in West Java and found that K-Means was better by the silhouette index. Internationally, spatial and space-time clustering of TB has been used to detect clusters at the district level in Pakistan (Ullah et al., 2020) and clusters of multidrug-resistant TB in South Africa (Mashamba et al., 2022). To the authors' knowledge, no study has compared K-Means and K-Medoids with three distances for TB in Central Java using indicators of cases, outcomes, and RO TB together.

The novelty of this study is therefore threefold: (1) a systematic comparison of two algorithms and three distances (six combinations) for the same TB data; (2) the use of ten indicators that include treatment outcomes and RO TB, and not only the number of cases; and (3) the translation of the best clustering into low, moderate, and high categories with profiles that can guide priorities for TB control in Central Java.

This study aims to compare the K-Means and K-Medoids algorithms in mapping the spread of TB in Central Java Province. The research questions are: (1) how many clusters are optimal for the data; (2) which combination of algorithm and distance gives the best cluster quality according to the Davies-Bouldin Index; and (3) what are the characteristics of the clusters formed and which districts/cities should be prioritized for TB control?

2. Method

This study uses secondary data obtained from two main sources, namely the Central Java Provincial Health Office and the Central Java Provincial BPS. The data used is data in 2024 which is the basis for building a dataset for clustering analysis in this study. This study uses the python programming language to process data. The initial data processing carried out is data preprocessing, namely data cleaning to be processed later by comparing the K-Means and K-Medoids algorithms with Euclidean, Manhattan, and Minkowski distances.

Each district/city was described by ten indicators: TB cases aged 0–14 years, TB cases aged over 15 years, cases in men, cases in women, cases cured, deaths, treatment failures, loss to follow-up, RO TB cases, and RO TB referral hospitals (X1 to X10 in Table 3). Because the indicators have different scales, they were standardized by z-scores before clustering. The number of clusters was chosen from the elbow and silhouette methods, each algorithm was run with each of the three distances, and the results were compared with the Davies-Bouldin Index.

2.1. Clustering

Cluster analysis classifies objects so that each object that has similar properties (with the highest level of similarity) will be grouped in the same cluster (Loklomin et al., 2022). Clustering is a grouping that uses unsupervised learning techniques, where there is no training phase and does not require labeling for each group (Priyatman et al., 2019). In general, clustering is divided into two types, namely hierarchical clustering and non-hierarchical clustering. Hierarchical clustering is a data grouping method that begins by combining two or more objects that have the highest level of similarity. In the hierarchical method, there are several types of cluster analysis used, such as single linkage, average linkage, and complete linkage. Meanwhile, non-hierarchical clustering is a data grouping method in which the number of clusters to be formed has been determined from the start. While the types of clustering analysis used in non-hierarchical methods, namely the K-Means, K-Medoids and Fuzzy C-Means methods.

2.2. K-Means

K-Means is a non-hierarchical clustering method used to group data into one or more clusters (Tendean & Purba, 2020). This method is known to be effective in handling large data (Rosida & Wijaya, 2023). The K-Means algorithm also has a good level of accuracy for object size, making it a fairly efficient and measurable algorithm in processing large amounts of data (Rosida & Wijaya, 2023).

The steps of the K-Means Clustering algorithm are as follows

1. Determine the value of k in the data set
2. Determining the center value (centroid). Determining the centroid value at the initial stage is done randomly, while at the iteration stage using the following formula.

$$V_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \quad (1)$$

In this formula, V_{ij} states the average centroid of Cluster i for variable j , N_i states the number of members of Cluster i , while i, k states the index of the Cluster, j states the index of the variable, and X_{kj} declare data value to k variable j for the Cluster

3. Calculate the distance between the centroid point and the object point using Euclidean distance, Manhattan distance and Minkowski distance..
4. Group objects based on distance to the nearest centroid.
5. Repeat steps 2 to 4, and iterate until the centroid reaches the optimal value.

2.3. K-Medoids

K-Medoids, also known as Partitioning Around Medoids (PAM), is a non-hierarchical method other than K-Means. This method is similar to K-Means, but uses medoids (the most representative center points) instead of centroids. Unlike K-Means, which is susceptible to the influence of outliers, the K-Medoids algorithm provides more stable and accurate clustering results in many real-world applications (Schubert & Rousseeuw, 2019).

There are several steps in the K-Medoids algorithm, namely:

1. Initialize k cluster centers.
2. Calculate each object to the nearest cluster using the distance measure equations Euclidean distance, Manhattan distance and Minkowski distance..
3. Randomly select objects from each cluster as new candidate medoids.
4. Calculate the distance of each object in each cluster with the new candidate medoids.
5. Calculate the total deviation (S) by taking the difference between the new total distance and the old total distance. If S is less than 0, then swap the objects with the data in the cluster to form a new set of k objects as medoids.
6. Repeat steps 3 to 5 until there are no changes in the medoids, so that you get the clusters and their respective cluster members.

2.4. Distance Selection

2.4.1. Euclidean distance

Euclidean distance is one of the distance measurement methods used to calculate the distance between two points in Euclidean space. This method is used to assess the level of similarity between data using the following Euclidean formula,

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

In this formula, d states the distance between x and y , mark x_i is the central cluster data to i whereas y_i is data on each data to i .

2.4.2. Manhattan distance

Manhattan distance is used to measure the absolute difference between the coordinates of two objects. This distance calculation is done using the following formula,

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

In this formula, d states the distance between x and y , mark x_i is the central cluster data to i whereas y_i is data on each data to i .

2.4.3. Minkowski distance

Minkowski distance is a metric in vector space where a norm is defined (normed vector space) and is considered as a generalization of Euclidean distance and Manhattan distance. The following is the formula for this method,

$$d(x, y) = \left(\sum_{i=1}^n (|x_i - y_i|^p) \right)^{\frac{1}{p}} \quad (4)$$

In this formula, d states the distance between x and y , mark x_i is the central cluster data to i whereas y_i is data on each data to i . Parameter p is a power value that determines the distance value.

2.5. Davies-Bouldin Index

Davies-Bouldin Index (DBI) was introduced in 1979 by David L. Davies and Donald W. Bouldin. The DBI evaluation method has the advantage in assessing the quality of clustering because it considers the cohesion and separation values produced. Cohesion refers to the proximity of data to the centroid of the cluster it follows, known as the sum of squares within cluster. Meanwhile, separation refers to the distance between centroids between clusters, known as the sum of squares between clusters. Based on the DBI calculation, the smaller the index value obtained (non-negative value ≥ 0), the better the quality of the resulting cluster is considered.

3. Results and Discussion

3.1. Application of K-Means Algorithm

3.1.1. Determining the Optimal k Value

Determining the optimal k value in the K-Means method is done using two approaches, namely the elbow and silhouette methods.

3.1.1.1. Elbow Method

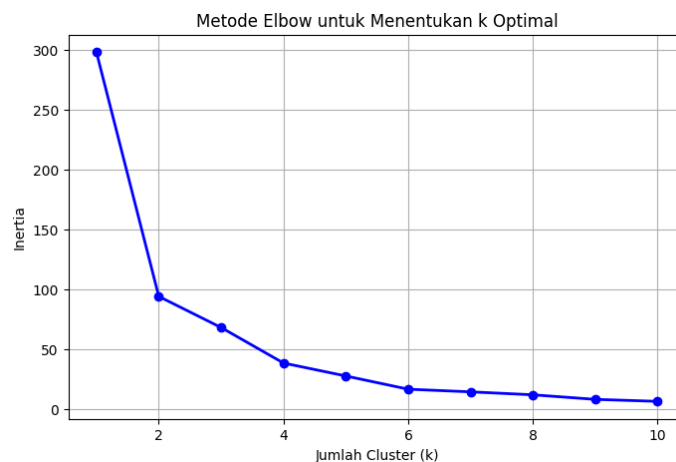


Figure 1. Elbow K-Means Method Graph

Based on Figure 1, the elbow graph shows that the Within-Cluster sum of squares (WCSS) value decreases as the number of clusters increases. However, the decrease begins to slow down after point $k=3$ or $k=4$, which forms an “elbow”. This point is considered the optimal number of clusters because after that, adding clusters does not provide a significant reduction in inertia.

3.1.1.2. Silhouette Method

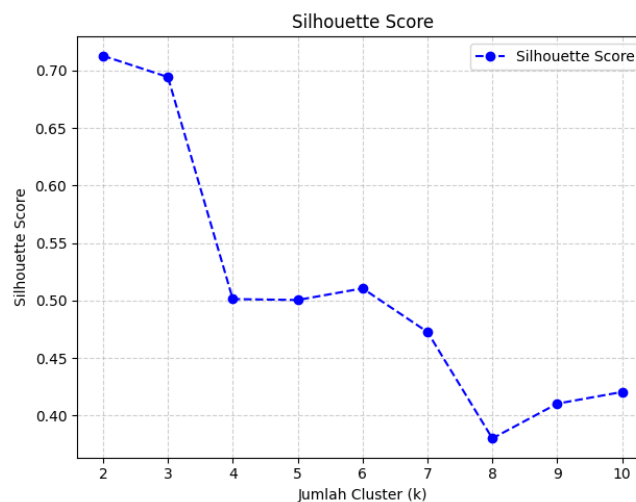


Figure 2. K-Means Silhouette Method Graph

Based on Figure 2 of the Silhouette Score graph above, the highest value occurs when the number of clusters is $k = 2$ with a score above 0.70. However, even though the score is high, the selection of $k = 2$ is considered less representative because it only produces two clusters that are too general and less spread out. Therefore, the number of clusters $k = 3$ is chosen as a more optimal alternative.

Based on the results of the two methods in determining the optimal k with the elbow method and the silhouette method, the optimal number of clusters obtained in the elbow method is 3 clusters, while the optimal cluster in the silhouette method is 2 clusters. From the two methods of determining the optimal k , $k = 3$ is chosen, because if $k = 2$ the clusters formed are less spread out..

3.1.2. Distance Selection

3.1.2.1. Euclidean distance

The grouping of districts/cities in Central Java Province based on the distribution of tuberculosis cases using this method is presented in the map visualization in Figure 3. Each color describes a different cluster group. Cluster 1 is depicted in yellow, cluster 2 is depicted in white, while cluster 3 is depicted in green.

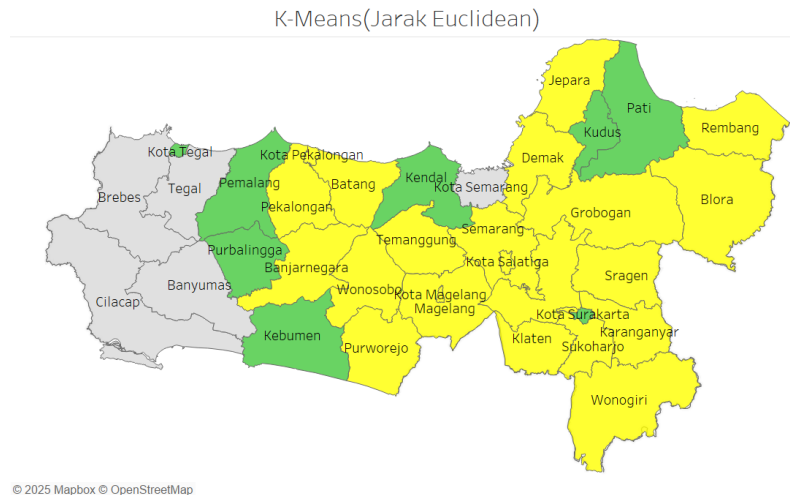


Figure 3. K-Means Cluster Map(Euclidean Distance)

3.1.2.2. Manhattan distance

The grouping of districts/cities in Central Java Province based on the distribution of tuberculosis cases using this method is presented in the map visualization in Figure 4. Each color describes a different cluster group. Cluster 1 is depicted in yellow, cluster 2 is depicted in white, while cluster 3 is depicted in green.

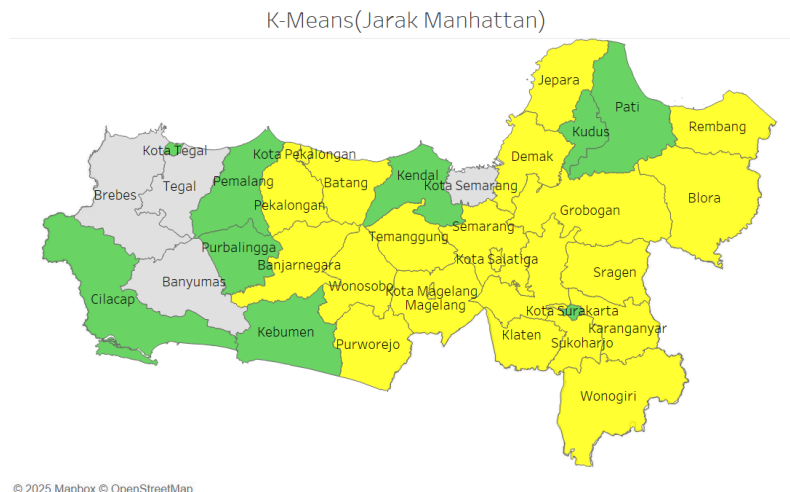


Figure 4. K-Means Cluster Map(Manhattan Distance)

3.1.2.3. Minkowski distance

The grouping of districts/cities in Central Java Province based on the distribution of tuberculosis cases using this method is presented in the map visualization in Figure 5. Each color describes a different cluster group. Cluster 1 is depicted in yellow, cluster 2 is depicted in white, while cluster 3 is depicted in green.



Figure 5. K-Means Cluster Map(Minkowski Distance)

3.2. Application of K-Medoids Algorithm

3.2.1. Determining the Optimal k Value

Determining the optimal k value in the K-Medoids method is done using two approaches, namely the elbow and silhouette methods.

3.2.1.1. Elbow Method

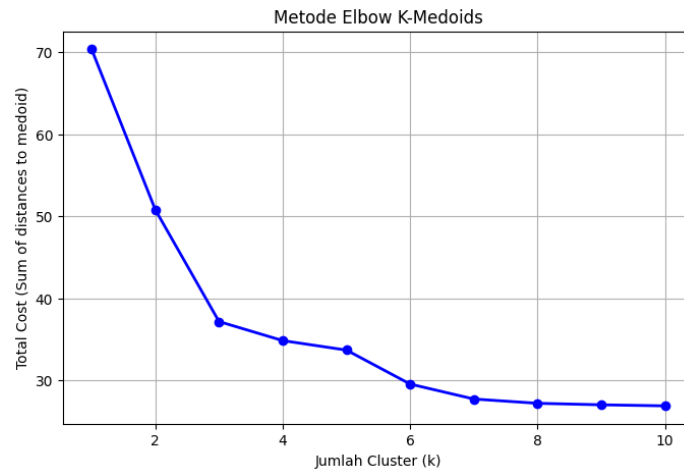


Figure 6. Elbow K-Medoids Method Graph

Based on Figure 6, the elbow method graph on the K-Medoids algorithm shows that the decrease in total cost (the number of distances to the medoid) begins to slow down significantly after the number of clusters $k = 3$. This point forms an "elbow" on the graph, which shows that increasing the number of clusters after $k = 3$ no longer provides a significant reduction in cost.

3.2.1.2. Silhouette Method

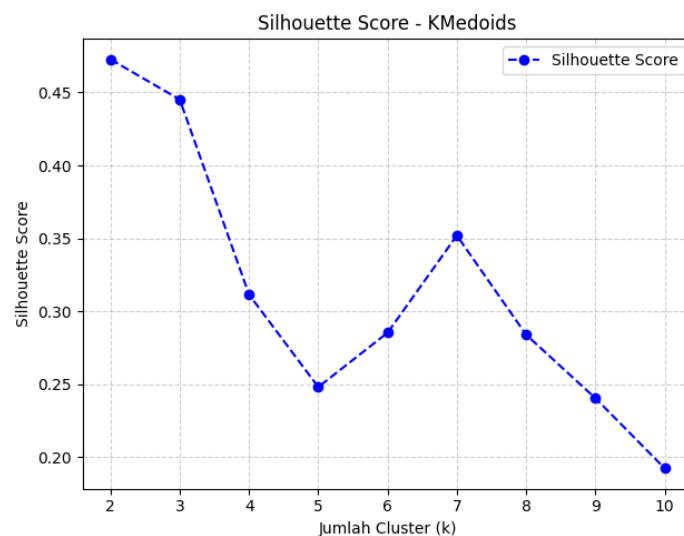


Figure 7. K-Medoids Silhouette Method Graph

Based on Figure 7 of the Silhouette Score graph above, the highest value occurs when the number of clusters is $k = 2$ with a score above 0.70. However, even though the score is high, the selection of $k = 2$ is considered less representative because it only produces two clusters that are too general and less spread out. Therefore, the number of clusters $k = 3$ is chosen as a more optimal alternative.

Based on the results of the two methods in determining the optimal k with the elbow method and the silhouette method, the optimal number of clusters obtained in the elbow method is 3 clusters, while the optimal cluster in the silhouette method is 2 clusters. From the two methods of determining the optimal k , $k = 3$ is chosen, because if $k = 2$ the clusters formed are less spread out.

3.2.2. Distance Selection

3.2.2.1. Euclidean distance

The grouping of districts/cities in Central Java Province based on the distribution of tuberculosis cases using this method is presented in the map visualization in Figure 8. Each color describes a different cluster group. Cluster 1 is depicted in yellow, cluster 2 is depicted in white, while cluster 3 is depicted in green.

K-Medoids(Jarak Euclidean)

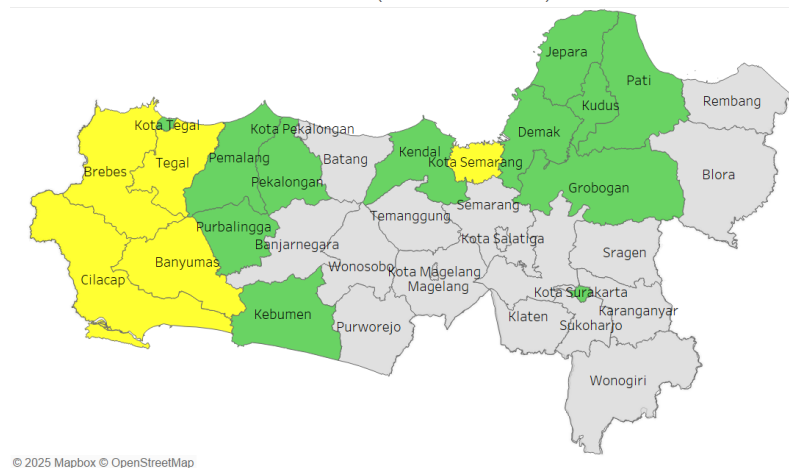


Figure 8. K-Medoids Cluster Map(Euclidean Distance)

3.2.2.2. Manhattan distance

The grouping of districts/cities in Central Java Province based on the distribution of tuberculosis cases using this method is presented in the map visualization in Figure 9. Each color describes a different cluster group. Cluster 1 is depicted in yellow, cluster 2 is depicted in white, while cluster 3 is depicted in green.

K-Medoids(Jarak Manhattan)

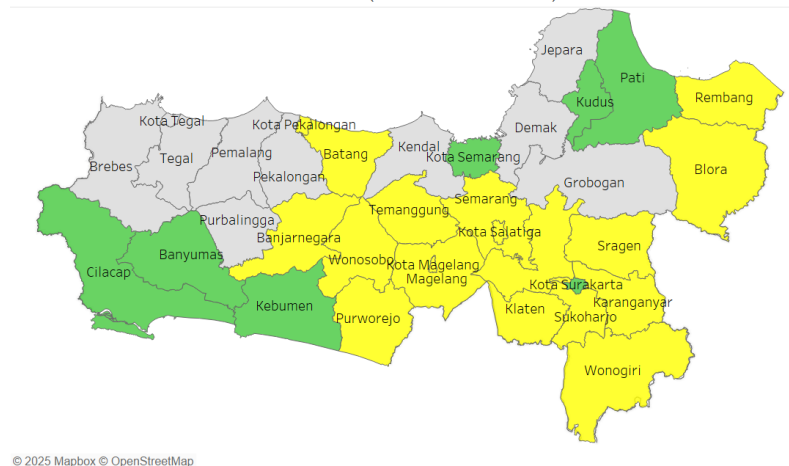


Figure 9. K-Medoids Cluster Map(Manhattan Distance)

3.2.2.3. Minkowski distance

The grouping of districts/cities in Central Java Province based on the distribution of tuberculosis cases using this method is presented in the map visualization in Figure 10. Each color describes a different cluster group. Cluster 1 is depicted in yellow, cluster 2 is depicted in white, while cluster 3 is depicted in green.

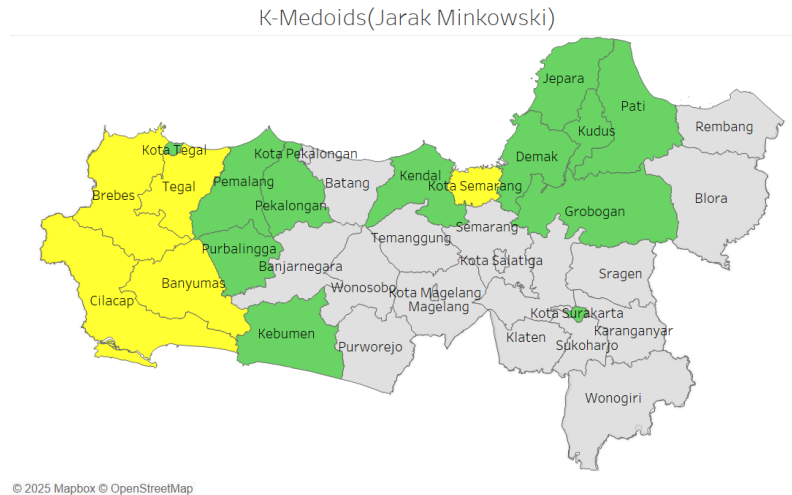


Figure 10. K-Medoids Cluster Map(Minkowski Distance)

3.3. Evaluation of the Best Method

The next step is to determine the method that produces the best cluster quality. Evaluation of cluster results in this study was carried out using the Davies-Bouldin Index (DBI) metric. A smaller DBI value (with a non-negative range ≥ 0) indicates that the cluster formed is better.

Table 1. Evaluation of the Best Method

Clustering Method		Validity Method
Distance		<i>Davies-Bouldin Index</i>
Algorithm		
<i>K-Means</i>	Euclidean	0.7822
	Manhattan	0.7789
	Minkowski	0.7822
<i>K-Medoids</i>	Euclidean	0.7803
	Manhattan	0.7950
	Minkowski	0.7803

Based on table 1, the K-Means algorithm with the Manhattan distance is better than other algorithms with the lowest DBI value of 0.7789. The findings of this study are in line with research (Utomo, 2021) which found that the K-Means algorithm is better than the K-Medoids algorithm because the DBI value of K-Means is lower than K-Medoids. Research (Febrian and Noviandi, 2024) showed that the K-Means algorithm with the Manhattan distance is better than the K-Means algorithm with the Euclidean distance. The clustering results using the Minkowski distance with parameter $p = 2$ show identical similarities to the clustering results using the Euclidean distance. This happens because mathematically, Minkowski with $p = 2$ is equivalent to Euclidean, resulting in the same distance calculation and cluster structure.

The DBI values in Table 1 differ very little: the best (K-Means with Manhattan distance, 0.7789) is only 0.4% lower than K-Means with Euclidean distance (0.7822) and 0.2% lower than K-Medoids with Euclidean distance (0.7803), and the worst (K-Medoids with Manhattan distance, 0.7950) is 2.1% higher. The five best combinations are therefore practically equal, and the choice of algorithm and distance has little effect on cluster quality in these data; the ranking could change with a different random initialization. The two algorithms also gave the same DBI for the Euclidean and Minkowski distances because the Minkowski distance with $p = 2$ is the Euclidean distance, so only two distances were effectively compared. Using other values of p would test the Minkowski distance.

Grouping of districts/cities in Central Java Province based on the distribution of tuberculosis cases using the K-Means method with Manhattan distance is presented in table 2, which shows the distribution of members in each cluster as follows.

Table 2. K-Means Cluster Results(Manhattan Distance)

Cluster	Regency/City	Number	Category
1	Banjarnegara, Batang, Blora, Boyolali, Demak, Grobogan, Jepara, Karanganyar, Klaten, Magelang, Pekalongan, Purworejo, Rembang, Semarang, Sragen, Sukoharjo, Temanggung, Wonogiri, Kab. Wonosobo, Kota Magelang, Kota Pekalongan, Kota Salatiga	22	Low
2	Banyumas, Brebes, Tegal, Kota Semarang	4	High

3	Kebumen, Cilacap, Kendal, Kudus, Pati, Pemalang, Purbalingga, Kota Surakarta, Kota Tegal	9	Moderate
---	--	---	----------

Determining the best clustering method is followed by the interpretation stage of the results and profiling of the formed clusters, which in this case uses the K-Means method with the Manhattan distance. The purpose of this stage is to identify the characteristics of each cluster. Profiling is done by analyzing the average value of each variable in each cluster. Details of the average values are presented in table 3.

Table 3. Cluster Average Value

Cluster	1	2	3
TB cases aged 0–14 years (X1)	-0.5074	2.3342	0.2030
TB cases aged over 15 years (X2)	-0.5716	2.3385	0.3578
TB cases in men (X3)	-0.5698	2.3784	0.3357
TB cases in women (X4)	-0.5755	2.4361	0.3241
Number of TB cases cured (X5)	-0.5721	2.3626	0.3485
Number of TB deaths (X6)	-0.4682	1.9909	0.2598
Number of TB treatment failures (X7)	-0.3566	0.6996	0.5608
Number of TB loss-to-follow-up cases (X8)	-0.4240	1.9067	0.1891
Drug-resistant (RO) TB cases (X9)	-0.4387	2.1133	0.1331
Referral hospitals for RO TB (X10)	-0.3525	1.4180	0.2312

The results of K-Means with Manhattan distance form three main clusters with the following member characteristics:

- Cluster 1 is included in the Low category, marked by a negative average value for all variables, meaning that districts/cities in this cluster have relatively low TB cases compared to other clusters. A negative Z-score value indicates that indicators such as TB cases aged 0–14 years (-0.5074), aged over 15 years (-0.5716), male cases (-0.5698), and female cases (-0.5755), are below the average of the 35 districts/cities. This indicates that the spread of TB in this region is still under control. In addition, the number of patients who recovered (-0.5721) was also low, reflecting a smaller volume of cases, not due to low treatment effectiveness. Severity indicators, such as the number of deaths (-0.4682), treatment failure (-0.3566), and loss to follow-up (-0.4240) also showed low values, meaning that the level of complications or success of treatment was relatively better. Therefore, Cluster 1 describes areas with a relatively safe and controlled TB situation, although prevention and tracing efforts still need to be maintained so that cases do not increase..
- Cluster 2 is categorized as High, with the largest z-score in all ten indicators and values above 1.9 in eight of them (for example 2.33 to 2.44 for the number of cases and 2.11 for RO TB cases). The areas in this cluster have a very high number of TB cases, especially among adults, RO TB, and deaths, and are a priority focus in controlling TB, in terms of service capacity, access to treatment, and case tracking. The areas categorized as high are Banyumas, Brebes, Tegal Regency, and Semarang City. The indicator of treatment failure is much less extreme in this cluster (0.70) than the indicators of the number of cases, which suggests that failure grows less than in proportion to the number of cases; this needs to be confirmed with rates.
- Cluster 3 is included in the Moderate category, with a Z-score value that is around the average (between 0.1 and 0.5). The areas in this Cluster have TB incidence rates that need to be watched out for, especially because they still show numbers that are not too low, although not as bad as Cluster 2. This Cluster can be a strategic area for further preventive interventions, to prevent a spike in cases that have the potential to enter the high category.

3.4. Discussion

The results show that the ranking of the methods is not decisive, because the DBI values are very close. The three clusters correspond to a gradient of TB burden on all indicators, with a low cluster of 22 districts/cities, a moderate cluster of 9, and a high cluster of 4. Because the indicators are numbers of cases and not rates per population, the high cluster probably reflects the size of the population as well as the burden of TB, and its districts should be checked against population data before they are read as the areas with the highest risk.

The finding that K-Means has a lower DBI than K-Medoids agrees with Utomo (2021) for Covid-19 data, and the advantage of Manhattan over Euclidean distance for K-Means agrees with Febrian and Noviandi (2024), but in this study the differences are small enough to be within the effect of initialization, so they are weaker than a clear superiority.

Compared with Faidah et al. (2025), who used K-Means and DBSCAN for TB in West Java and found that K-Means was better by the silhouette index, the present study adds K-Medoids, three distances, and outcome and RO TB indicators. Loklomin et al. (2022) clustered TB by determinants, whereas this study clusters by the burden of cases and outcomes, which gives a complementary view for planning. Because the data sets and validity indices differ, the comparison is limited to the direction of the findings.

The novelty of the findings is that the best of six combinations of algorithm and distance gives clusters with a clear profile, that the choice among the combinations has only a small effect on cluster quality, and that the high cluster stands out in all ten indicators, including RO TB and treatment outcomes, and not only in the number of cases.

These findings imply the following recommendations for TB control in Central Java. (1) For the high cluster (Banyumas, Brebes, Tegal Regency, and Semarang City): increase the capacity of health services, expand access to treatment, strengthen case tracking and early detection, and reinforce RO TB services, because the indicator of RO TB referral hospitals is high (1.42) in this cluster. (2) For the moderate cluster (Kebumen, Cilacap, Kendal, Kudus, Pati, Pemalang, Purbalingga, Surakarta City, and Tegal City): keep a watch list and carry out preventive interventions to prevent a spike in cases. (3) For the low cluster: maintain surveillance and public education on TB prevention. (4) For analysis: use K-Means with Manhattan distance as the working method but verify the stability of the clusters, and repeat the clustering with incidence rates before allocating resources.

Several limitations should be noted. The indicators are numbers and not rates, only one year (2024) and 35 objects were analyzed, the number of clusters was chosen by judgment because the silhouette method favors two clusters, the differences in the DBI are small, and the clustering does not use the geographic position of the districts, so that neighboring districts are not necessarily in the same cluster. Future research should use incidence rates, several years, other validity indices, and spatial methods.

4. Conclusion

The three research questions are answered as follows. First, the optimal number of clusters is three: the elbow method points to $k = 3$, and $k = 3$ was chosen over the $k = 2$ favored by the silhouette method because two clusters are too general. Second, the K-Means method with Manhattan distance gives the best cluster quality (DBI of 0.7789), against 0.7822 for K-Means with Euclidean and Minkowski distances and 0.7803, 0.7950, and 0.7803 for K-Medoids with Euclidean, Manhattan, and Minkowski distances; the differences are small, so the five best combinations are practically equal. Third, the K-Means clusters with Manhattan distance are a low cluster of 22 districts/cities, a moderate cluster of 9, and a high cluster of 4 (Banyumas, Brebes, Tegal Regency, and Semarang City), the last having z-scores above 1.9 in eight of ten indicators.

As a follow-up, the government and health policy makers are recommended to give special attention to the high cluster by increasing the capacity of health services, expanding access to treatment, increasing public education on TB prevention, and optimizing case tracking and early detection, to keep a watch list for the moderate cluster, and to repeat the clustering with incidence rates and several years before allocating resources.

References

- Alam, M. M., Khan, M. K., & Islam, N. (2019). *An overview of epidemiology of tuberculosis*.
- Dani, A. T. R., Wahyuningsih, S., & Rizki, N. A. (2019). Penerapan hierarchical clustering metode agglomerative pada data runtun waktu. *Jambura Journal of Mathematics*, 1(2), 64–78. <https://doi.org/10.34312/jiom.v1i2.2354>
- Faidah, D. Y., Destin, D., Anggina, F. A., & Caesar, M. I. (2025). Assessing the performance of K-means and DBSCAN clustering methods in tuberculosis mapping. *Communications in Mathematical Biology and Neuroscience*, 2025, Article 15.
- Febrian, N., & Novianti, N. (2024). Perbandingan Manhattan dan Euclidean distance untuk pengelompokan penyakit jantung menggunakan algoritma K-Means. *ICIT Journal*, 10(1), 61–70. <https://doi.org/10.33050/icit.v10i1.2860>
- Kementerian Kesehatan Republik Indonesia. (2023). *Profil kesehatan Indonesia 2022*. Kementerian Kesehatan Republik Indonesia.
- Loklomin, S. B., Matdoan, M. Y., & Lema, H. (2022). Clustering determinan tuberculosis kabupaten/kota di Provinsi Maluku Utara. *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, 10(2), 9–16. <https://doi.org/10.24252/msa.v10i2.27852>
- Mashamba, M. A., Tanser, F., Afagbedzi, S., & Beke, A. (2022). Multi-drug-resistant tuberculosis clusters in Mpumalanga province, South Africa, 2013–2016: A spatial analysis. *Tropical Medicine & International Health*, 27(2), 185–191.

- Priyatman, H., Sajid, F., & Haldivany, D. (2019). Klasterisasi menggunakan algoritma K-Means clustering untuk memprediksi waktu kelulusan mahasiswa. *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, 5(1), 62. <https://doi.org/10.26418/jp.v5i1.29611>
- Rosida, W., & Wijaya, Y. A. (2023). Klasterisasi penyakit HIV/AIDS di Jawa Barat menggunakan algoritma K-Means clustering. *Blend Sains Jurnal Teknik*, 1(4), 306–315. <https://doi.org/10.56211/blendsains.v1i4.235>
- Schubert, E., & Rousseeuw, P. J. (2019). Faster k-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms. In *Similarity Search and Applications (SISAP 2019)* (pp. 171–187). Springer. https://doi.org/10.1007/978-3-030-32047-8_16
- Syamsu, Muhajir, & Widodo, W. (2021). Peran data science dan data scientist untuk mentransformasi data dalam industri 4.0. *Jurnal Teknologi Informasi (JUTECH)*, 2(1), 27–36. <https://doi.org/10.32546/jutech.v2i1.1540>
- Tendean, T., & Purba, W. (2020). Analisis cluster provinsi Indonesia berdasarkan produksi bahan pangan menggunakan algoritma K-Means. *Jurnal Sains dan Teknologi*, 1(2), 5–11.
- Ullah, S., Daud, H., Dass, S. C., & Fanaee-T, H. (2020). Space-time clustering characteristics of tuberculosis in Khyber Pakhtunkhwa Province, Pakistan, 2015–2019. *International Journal of Environmental Research and Public Health*, 17(4).
- Utomo, W. (2021). The comparison of K-Means and K-Medoids algorithms for clustering the spread of the Covid-19 outbreak in Indonesia. *ILKOM Jurnal Ilmiah*, 13(1), 31–35. <https://doi.org/10.33096/ilkom.v13i1.763.31-35>
- World Health Organization. (2024). *Global tuberculosis report 2024*. World Health Organization. <https://who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024>