

Application of Spatial K-Means Clustering for Identifying Areas at High Risk of Tuberculosis (TB) in Indonesia in 2025

Tisa Aulia¹, Ilham Hayadi²

^{1,2}Universitas Bengkulu, Indonesia

Email: dyanikovigetta@gmail.com

Abstract

This study aims to classify Indonesian provinces based on tuberculosis (TB) characteristics projected for 2025 using the K-Means Clustering method. The analysis covered 38 provinces using two indicators: the TB Case Detection Rate (JPTBC) and the TB Treatment Success Rate (JKPTBC). Data were standardized using the Z-score method prior to clustering. The optimal number of clusters was determined using the Silhouette Coefficient, resulting in seven distinct clusters for analysis. The findings reveal variations in TB characteristics across provinces, as evidenced by the spatial distribution of the seven clusters. The mapping indicates that TB conditions in Indonesia are heterogeneous, necessitating TB control strategies tailored to the specific characteristics of each provincial group. These clustering results can serve as a basis for prioritizing interventions and formulating TB control policies in Indonesia.

Keywords: TB, K-Means Clustering, Spatial, JPTBC, JKPTBC, Indonesia.

A. INTRODUCTION

Tuberculosis (TB) is an infectious disease that remains a major public health concern because it is transmitted through the air and can spread in both household and community settings (World Health Organization (WHO) (Shaweno et al., 2018). The WHO estimated that approximately 10.7 million people developed TB in 2024, with around 1.23 million deaths attributed to the disease (WHO, 2025). Indonesia is among the countries with the highest TB burden worldwide and accounts for approximately 10% of global TB cases. Therefore, TB control in Indonesia requires data-driven and area-based approaches (Kementerian Kesehatan RI, 2024).

The distribution of TB cases in Indonesia is not uniform across regions (Lönnroth et al., 2009; Baker et al., 2008; Harling & Castro, 2014). Studies on the spatial epidemiology of TB have shown that TB cases can form geographic patterns of clustering (spatial clustering), in which areas with high numbers of cases may be adjacent to other areas with similar characteristics (Munch et al., 2003; Shaweno et al., 2018; Tiwari et al., 2006). Spatial analysis is therefore necessary to identify areas with similar TB case characteristics so that health interventions can be targeted more effectively (Teibo et al., 2023; Onozaki et al., 2015; WHO, 2023).

One method that can be used to group regions based on similarities in characteristics is K-Means Clustering. K-Means Clustering is an unsupervised learning method that partitions objects into a predetermined number of groups based on the distance or degree of similarity among observations (MacQueen, 1967; Hartigan & Wong, 1979; Jain, 2010). The application of K-Means in TB analysis has

been conducted in South Sulawesi, demonstrating that the method can classify regions based on the number of TB cases and identify areas with relatively high numbers of cases (Mar'ah *et al.*, 2025).

The application of Spatial K-Means Clustering can be used to identify groups of provinces in Indonesia based on the characteristics of TB cases in 2025, thereby providing an overview of regions with different levels of TB vulnerability. This clustering is expected to provide information on priority areas with high TB case characteristics and support the development of area-based disease prevention and control strategies (WHO, 2025; Faidah *et al.*, 2025; Mar'ah *et al.*, 2025). Therefore, this study applies Spatial K-Means Clustering to identify TB-vulnerable areas in Indonesia in 2025 as an effort to apply spatial statistical methods to support TB mapping and control in Indonesia.

B. METHOD

This study uses a quantitative approach to analyze TB cases in Indonesia in 2025. This study uses secondary data from the Ministry of Health of the Republic of Indonesia and the Central Statistics Agency (BPS). The research data includes the number of TB cases and the TB treatment success rate in each province in Indonesia.

The researcher standardizes the data using the Z-score method before the clustering process. The researcher determines the optimal number of clusters using the Silhouette Coefficient. The researcher groups provinces in Indonesia using the K-Means Clustering method based on the similarity of TB case characteristics.

The researcher evaluates the clustering results using the Silhouette Coefficient. The researcher interprets each cluster based on the characteristics of the TB indicators. The researcher presents the clustering results in the form of a map to identify areas with high, medium, and low levels of TB vulnerability.

C. RESULTS AND DISCUSSION

1. Data Standardization

Data standardization is the process of changing the scale of numerical values from several variables into the same range or distribution without changing the patterns or relationships between variables (Han *et al.*, 2012). The standardization used is Z-Score. The Z-Score formula used is: $Z = (X_i - \mu) / \sigma$ used for the transformation of numerical data that changes the values of variables so that they have a mean of 0 and a standard deviation of 1 (Jain, 2010). Standardization prevents variables with larger units or scales from dominating the calculations and ensures that each indicator contributes equally to the determination of clusters (Milligan & Cooper, 1988). The results of the mean and standard deviation values of the data on the number of TB cases and the TB treatment success rate in each province in Indonesia are as follows:

Table 1. Results of the mean and standard deviation of the data on the number of TB cases and the TB treatment success rate in each province in Indonesia

μ		σ	
JPTBC	JKPTBC	JPTBC	JKPTBC
71.60526	81.36842	13.72222	7.091758

Table 1 shows the mean and standard deviation values of the data on the number of TB cases and the TB treatment success rate in each province in Indonesia. The mean value shows that nationally, the average TB case detection rate (JPTBC) was 71.61 and the average TB treatment success rate (JKPTBC) was 81.37%. The standard deviation values indicate a data distribution level of 13.72 for JPTBC and 7.09 for JKPTBC, indicating that the distribution of treatment success rates among provinces is relatively more homogeneous compared to case detection. These two values serve as the subtracting and dividing values in the formula to align the scales of the two indicators, so that each variable has an equal contribution and prevents variables with larger numerical scales from dominating cluster formation in the K-Means algorithm. The results of the data standardization are as follows:

Table 2. Data Standardization Table

Province	JPTBC	JKPTBC
Aceh	-0.26273169	-0.0519506
North Sumatra	0.17451521	-0.33396812
West Sumatra	-0.6999786	0.65309321
Riau	0.02876624	0.65309321
Jambi	-0.48135515	0.65309321
South Sumatra	-0.62710412	1.0761195
Bengkulu	-1.72022139	-0.75699441
Lampung	-0.48135515	1.49914579
Bangka Belitung Islands	-0.62710412	0.51208445
Riau Islands	-1.57447242	0.08905817
DKI Jakarta	1.04900903	-0.47497688
West Java	1.85062836	0.65309321
Central Java	0.90326006	0.65309321
DI Yogyakarta	0.2473897	0.51208445
East Java	0.39313867	0.93511074
Banten	2.94374563	1.35813702
Bali	0.83038557	-0.0519506
West Nusa Tenggara	-0.84572757	0.79410198
East Nusa Tenggara	-0.6999786	0.51208445
West Kalimantan	0.2473897	0.65309321
Central Kalimantan	0.39313867	-0.61598564
South Kalimantan	0.39313867	-0.47497688
East Kalimantan	-1.42872345	0.23006693
North Kalimantan	-0.55422963	-1.74405574

North Sulawesi	-0.33560618	0.08905817
Central Sulawesi	0.90326006	-0.19295936
South Sulawesi	-0.48135515	0.37107569
Southeast Sulawesi	0.02876624	-0.0519506
Gorontalo	0.39313867	0.93511074
West Sulawesi	0.61176212	0.79410198
Maluku	-0.62710412	0.51208445
North Maluku	0.61176212	-0.89800317
West Papua	1.194758	-1.18002069
Southwest Papua	0.75751109	-2.02607327
Papua	-0.99147654	-0.89800317
South Papua	0.68463661	-0.19295936
Central Papua	0.02876624	-0.61598564
Highland Papua	-2.23034278	-3.57716965

Table 2 presents the results of the Z-score data standardization, showing variations in TB characteristics among provinces in Indonesia relative to the national average, with areas such as Banten and West Java classified as having optimal performance because they recorded case detection rates and recovery rates above the average. Provinces such as DKI Jakarta, West Papua, and Southwest Papua face the challenge of high case detection that has not yet been accompanied by adequate recovery rates. Areas such as Lampung, South Sumatra, and NTB recorded high treatment success despite their case detection remaining below the national average, while areas such as Highland Papua, Papua, and North Kalimantan were identified as the most vulnerable due to low case detection as well as patient recovery rates, thus requiring priority program interventions and comprehensive strengthening of health facilities.

2. Silhouette Coefficient

Silhouette Coefficient is a metric used to measure the quality and accuracy of clustering results in a dataset (Rousseeuw, 1987). This metric assesses how similar an object is to its own cluster compared to its nearest neighboring cluster. The formula for the Silhouette Coefficient is as follows: $s(i) = (b(i) - a(i)) / (\max(b(i), a(i)))$.

The value of the Silhouette Coefficient formula indicates the clustering of the dataset. A Silhouette Coefficient value greater than 1 indicates that the distance between objects within the cluster is very close, while the distance to other clusters is very far. A Silhouette Coefficient value approaching 0 indicates that the object is located at the boundary between two clusters, so the clustering is less distinct. A Silhouette Coefficient value approaching -1 indicates that the object is likely assigned to the wrong cluster because it is more similar to objects in the neighboring cluster. The results of the Silhouette Coefficient are as follows:

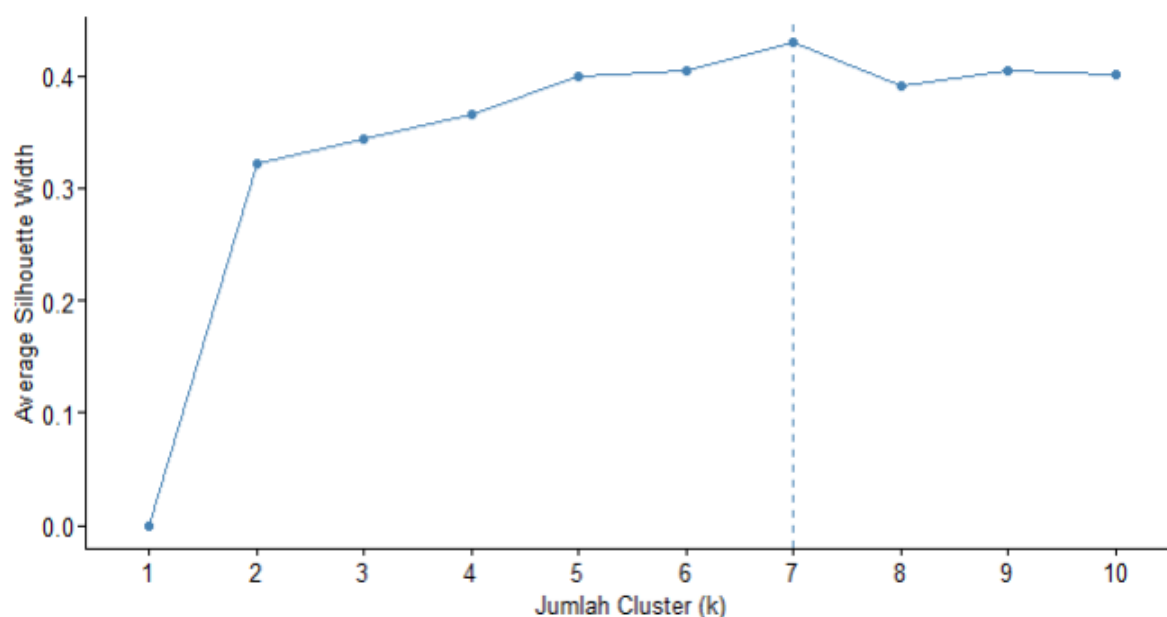


Figure 1. Silhouette Coefficient

Figure 1 shows the Silhouette Coefficient graph for determining the optimal K. The Average Silhouette Width value is calculated for various numbers of clusters, $K = 1$ to $K = 10$, to evaluate the quality of the clustering structure. The graph shows that the average silhouette value increases significantly from $K = 1$ until reaching its peak at $K = 7$, with a value of approximately 0.43. The dashed line at $K = 7$ indicates that, mathematically, dividing the data into 7 clusters provides the most optimal level of separation between clusters and density within the clusters.

3. K-Means Clustering

K-Means Clustering is one of the unsupervised learning methods used to group a number of objects into several groups based on the similarity of their characteristics. This method divides the data into a number of K clusters, with each object assigned to the cluster with the nearest center. The clustering process is carried out by determining the initial centers, allocating each object to a cluster based on the nearest distance, and then updating the centroid positions based on the average of the members in each cluster. This process is performed iteratively until a relatively stable cluster division is obtained (MacQueen, 1967; Hartigan & Wong, 1979). K-Means Clustering has a Silhouette Score in determining its clusters. Based on the Silhouette Score, there are 38 units divided into 7 clusters. The Average Silhouette Width value indicates how well each unit in a cluster fits its own cluster compared to other clusters. The clustering results based on the Silhouette Score are as follows:

Table 3. Silhouette Score Values in the Clusters

Cluster	Number of Members	Average Silhouette
1	2	0,3053
2	7	0,5847
3	2	0,1882
4	1	0,0000

5	11	0,4849
6	11	0,4105
7	4	0,3554

Table 3 presents the results of the Silhouette Coefficient analysis. The results of the Silhouette Coefficient analysis show that there are 38 observation units divided into 7 clusters with different clustering qualities. Cluster 2 has the highest average silhouette value of 0.5847 with 7 members, indicating the best clustering. Clusters 5 and 6 also show fairly good separation. Clusters 7 and 1 have moderate separation, while Cluster 3 has the lowest value of 0.1882, indicating a possible similarity with other clusters. Cluster 4 has only 1 member with a silhouette value of 0.0000, so its separation quality cannot be optimally assessed. These results show that the formation of 7 clusters is sufficiently capable of grouping the data based on its characteristics, although there are several clusters with relatively weak separation.

The summary results of the Silhouette Score values are as follows:

Table 4. Summary of Silhouette Score Values

Statistic	Value
Minimum	0,0000
First Quartile	0,2904
Median	0,4657
Average	0,4303
Third Quartile	0,5848
Maximum	0,6973

Table 4 shows that the Silhouette Coefficient has an average value of 0.4303 with a median value of 0.4657, indicating that overall, the grouping of the data into clusters is fairly good. The minimum value of 0.0000 indicates that there is a unit with an unclear position within its cluster, while the maximum value of 0.6973 indicates that there is a unit that is highly suitable for the cluster in which it is located. In addition, 50% of the data have silhouette values between 0.2904 and 0.584; thus, these results indicate that most units have been grouped fairly appropriately, although there are still several units with relatively low levels of separation between clusters. After that, the clustering results for each province were obtained, as follows:

Table 5. Cluster of Each Province

Province	Cluster	JPTBC	JKPTBC
West Java	1	97	86
Banten	1	112	91
Riau	2	72	86
Central Java	2	84	86
DI Yogyakarta	2	75	85
East Java	2	77	88
West Kalimantan	2	75	86
Gorontalo	2	77	88
West Sulawesi	2	80	87

North Kalimantan	3	64	69
Southwest Papua	3	82	67
Highland Papua	4	41	56
Aceh	5	68	81
West Sumatra	5	62	86
Jambi	5	65	86
South Sumatra	5	63	89
Lampung	5	65	92
Bangka Belitung Islands	5	63	85
West Nusa Tenggara	5	60	87
East Nusa Tenggara	5	62	85
North Sulawesi	5	67	82
South Sulawesi	5	65	84
Maluku	5	63	85
North Sumatra	6	74	79
DKI Jakarta	6	86	78
Bali	6	83	81
Central Kalimantan	6	77	77
South Kalimantan	6	77	78
Central Sulawesi	6	84	80
Southeast Sulawesi	6	72	81
North Maluku	6	80	75
West Papua	6	88	73
South Papua	6	81	80
Central Papua	6	72	77
Bengkulu	7	48	76
Riau Islands	7	50	82
East Kalimantan	7	52	83
Papua	7	58	75

Table 4 shows that 38 provinces are divided into 7 clusters based on the JPTBC and JKPTBC indicators. Cluster 2 consists of 7 provinces with relatively low to moderate JPTBC values and relatively high JKPTBC, while Cluster 5 is the largest cluster with 11 provinces that have relatively low JPTBC and fairly high JKPTBC. Cluster 6 consists of 11 provinces with relatively higher JPTBC but tends to have lower JKPTBC. Meanwhile, Cluster 1 only includes West Java and Banten with relatively high JPTBC, while Cluster 3 consists of North Kalimantan and Southwest Papua. Cluster 4 consists only of Highland Papua with the lowest JPTBC and JKPTBC values. Cluster 7 includes 4 provinces with the lowest JPTBC, but relatively high JKPTBC. In general, this clustering shows differences in characteristics among provinces based on the levels of JPTBC and JKPTBC. The visualization results of the clustering of each province in Indonesia are as follows:



Figure 2. Visualization of K-Means Clusters for Each Province in Indonesia

Figure 2 is a visualization of the K-Means Clusters of TB in Indonesia in 2025, showing 7 groups of provinces with different characteristics based on their positions on the JPTBC and JKPTBC dimensions. Cluster 1 appears to be the most separated, particularly Banten and West Java, which show TB characteristics that are quite different compared to other provinces. Cluster 2 consists mainly of provinces on the island of Java and several areas in Kalimantan or Sulawesi, such as East Java, Central Java, DI Yogyakarta, Riau, and West Kalimantan, which have relatively high similarity in characteristics. Cluster 3 includes a number of provinces in eastern and central Indonesia, while Cluster 4 is dominated by provinces such as North Kalimantan and Southwest Papua, which are located relatively far in the negative direction. Cluster 5 consists largely of provinces in the Sumatra region and parts of Sulawesi, while Cluster 6 includes Bali, Nusa Tenggara, Maluku, Papua, and several other provinces. Cluster 7 appears to be dominated by areas such as East Kalimantan, Riau Islands, Bengkulu, and Papua. Overall, the graph shows that provinces located close to each other have more similar TB characteristics, while provinces that are located far apart or are outliers, such as Highland Papua, Banten, and Southwest Papua, have characteristics that are the most different from those of other groups of provinces.

4. Spatial K-Means Clustering

Spatial K-Means Clustering is a development of the K-Means Clustering method used to group objects based on similarities in attribute characteristics while also considering location or geographical coordinates. K-Means is used for grouping based on the distance between observations and the centralized mean, while in Spatial K-Means, spatial information, such as geographical coordinates, can be included as part of the variables or assigned a certain weight in the clustering process. This method can produce groups of regions that have relatively similar characteristics and are geographically close to each other. Wang et al. (2024) explains that Spatial K-

Means is an adaptation of K-Means that treats the coordinates of each spatial unit as attributes so that similarities in characteristics and location can be considered simultaneously. Regional analysis is used in Spatial K-Means to identify groups of areas that have similar patterns or characteristics, for example, in disease mapping, socio-economic conditions, and environmental characteristics. Each object or region will be assigned to a cluster based on its proximity to the center of the group, so that objects within one cluster have more similar characteristics compared to objects in other clusters. This approach also allows researchers to observe the geographical clustering patterns of a phenomenon through the mapping of cluster results. The results of Spatial K-Means Clustering for TB cases with the indicators of TB case detection rate (JPTBC) and TB treatment success rate (JKPTBC) are as follows:



Figure 3. Mapping of K-Means Clusters for Each Province in Indonesia

Figure 3 presents a mapping of the K-Means clustering results, showing that the 38 provinces in Indonesia are divided into seven clusters based on the characteristics of JPTBC and JKPTBC. Spatially, each color represents a group of provinces with relatively similar characteristics based on these two indicators. The distribution of the clusters does not completely follow geographical boundaries, as provinces located on the same island may belong to different clusters. Cluster 1 is observed in several areas, including West Java, while Cluster 2 is distributed across several provinces in Sumatra, Java, Kalimantan, and Sulawesi. Cluster 3 is found in several areas such as North Kalimantan and several other provinces, while Cluster 4 appears in several provinces in the Sulawesi region and eastern Indonesia. Clusters 5 and 6 also have relatively wide distributions, covering provinces in Sumatra, Kalimantan, Sulawesi, Bali, Nusa Tenggara, Maluku, and Papua. Meanwhile, Cluster 7 is represented by the pink color and appears prominently in the Papua region. Overall, the map shows heterogeneity in TB characteristics among provinces in Indonesia, indicating that the level of TB vulnerability cannot be explained solely based on geographical proximity. This clustering provides an overview that each cluster has different JPTBC and JKPTBC characteristics, which can serve as a basis for determining priorities and TB control strategies that are more specific according to the characteristics of each group of provinces.

D. CONCLUSION

Based on the results of the K-Means Clustering analysis of 38 provinces in Indonesia in 2025 using the JPTBC (TB Case Detection Rate) and JKPTBC (TB Treatment Success Rate) indicators, it can be concluded that provinces in Indonesia have diverse characteristics of TB prevention and control. The clustering results with 7 clusters show that provinces with relatively similar JPTBC and JKPTBC characteristics tend to be in the same group, although their distribution is not always based on geographical proximity. The spatial mapping also shows heterogeneity among provinces, with several provinces having characteristics that differ considerably from those of other provinces. This indicates that the TB conditions in Indonesia are not homogeneous, so TB control and prevention strategies need to be adjusted to the characteristics of each cluster. The clustering results can be used as a basis for identifying groups of provinces with relatively similar conditions and can help the government determine intervention priorities, resource allocation, and more targeted TB control policies based on regional characteristics.

ACKNOWLEDGMENTS

The researcher would like to express gratitude to God Almighty, the supervisors, and all parties who have contributed through their cooperation and valuable contributions to the preparation of this journal, as well as those who have been willing to take the time to provide valuable data and information for this research.

REFERENCES

1. Baker, M., Das, D., Venugopal, K., & Howden-Chapman, P. (2008). Tuberculosis associated with household crowding in a developed country. *Journal of Epidemiology and Community Health*, 62(8), 715–721. <https://doi.org/10.1136/jech.2007.063610>
2. 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. <https://doi.org/10.28919/cmbn/9039>
3. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Morgan Kaufmann.
4. Harling, G., & Castro, M. C. (2014). A spatial analysis of social and economic determinants of tuberculosis in Brazil. *Health & Place*, 25, 56–67. <https://doi.org/10.1016/j.healthplace.2013.10.008>
5. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, 28(1), 100–108. <https://doi.org/10.2307/2346830>
6. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
7. Kementerian Kesehatan Republik Indonesia. (2024). *Profil kesehatan Indonesia 2023*. Kementerian Kesehatan Republik Indonesia.
8. Lönnroth, K., Jaramillo, E., Williams, B. G., Dye, C., & Raviglione, M. (2009).

- Drivers of tuberculosis epidemics: The role of risk factors and social determinants. *Social Science & Medicine*, 68(12), 2240–2246. <https://doi.org/10.1016/j.socscimed.2009.03.041>
9. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.
 10. Mar'ah, Z., Hafid, H., & Meliyana R, S. M. (2025). Epidemiological mapping of Tuberculosis in South Sulawesi Using Local Indicators of Spatial Association (Lisa) And K-Means clustering. *Sainsmat: Jurnal Ilmiah Ilmu Pengetahuan Alam*, 14(01), 1–11. <https://doi.org/10.35580/sainsmat141665022025>
 11. Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5(2), 181–204. <https://doi.org/10.1007/BF01897163>
 12. Munch, Z., Van Lill, S. W. P., Booysen, C. N., Zietsman, H. L., Enarson, D. A., & Beyers, N. (2003). Tuberculosis transmission patterns in a high-incidence area: A spatial analysis. *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease*, 7(3), 271–277.
 13. Onozaki, I., Law, I., Sismanidis, C., Zignol, M., & Glaziou, P. (2015). National tuberculosis prevalence surveys in Asia, 1990–2012: An overview of results and lessons learned. *Tropical Medicine & International Health*, 20(9), 1128–1145. <https://doi.org/10.1111/tmi.12534>
 14. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
 15. Shaweno, D., Karmakar, M., Alene, K. A., Ragonnet, R., Clements, A. C., Trauer, J. M., Denholm, J. T., & McBryde, E. S. (2018). Methods used in the spatial analysis of tuberculosis epidemiology: A systematic review. *BMC Medicine*, 16(1), 193. <https://doi.org/10.1186/s12916-018-1178-4>
 16. Teibo, T. K. A., Andrade, R. L. de P., Rosa, R. J., Tavares, R. B. V., Berra, T. Z., & Arcêncio, R. A. (2023). Geo-spatial high-risk clusters of Tuberculosis in the global general population: A systematic review. *BMC Public Health*, 23(1), 1586. <https://doi.org/10.1186/s12889-023-16493-y>
 17. Tiwari, N., Adhikari, C. M. S., Tewari, A., & Kandpal, V. (2006). Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic. *International Journal of Health Geographics*, 5, 33. <https://doi.org/10.1186/1476-072X-5-33>
 18. Wang, H., Song, C., Wang, J., & Gao, P. (2024). A raster-based spatial clustering method with robustness to spatial outliers. *Scientific Reports*, 14, 4103.
 19. World Health Organization. (2023). *Global tuberculosis report 2023*. World Health Organization.
 20. World Health Organization. (2024). *Global tuberculosis report 2024*. World Health Organization.

21. World Health Organization. (2025). *Global tuberculosis report 2025*. World Health Organization. <https://doi.org/10.2471/9789240116924>