Clustering Of Regions With Potential For A Tsunami In Indonesia Using The DBSCAN Method (Data Study for 1822 – 2022)

To cite this article: Avisena and Melany Febrina 2024 *J. Phys.: Conf. Ser.* **2734** 012047

Clustering Of Regions With Potential For A Tsunami In Indonesia Using The DBSCAN Method (Data Study for 1822 – 2022)

Avisena, Melany Febrina

Physics Department, Institut Teknologi Sumatera, Lampung, Indonesia

E-mail: avisena.119110015@student.itera.ac.id, Melany.febrina@fi.itera.ac.id

Abstract. Indonesia is a country comprising many islands and having an extensive coastline where coastal communities frequently engage in various activities. Tsunamis are a natural disaster risk in these coastal regions. This study aims to identify areas prone to tsunamis and analyze their characteristics using variables such as longitude, latitude, focal depth, and earthquake magnitude. The Density-Based Spatial Clustering of Application with Noise (DBSCAN) and OPTICS algorithms were used to group the tsunami datasets.

1. Introduction

Indonesia is widely recognized as an archipelagic nation with a lengthy coastline. The intensity of the sea reaches 70 percent of the total area, making the abundance of natural resources very large. Many Indonesians in coastal areas still take advantage of the natural wealth of the sea. This is a reason for them to continue their activities in the coastal area. The sea is their only source of economic and social life [1].

Common activities in coastal areas besides living are tourism locations. Tourism is part of social and economic activities that can be enjoyed and is a source of income for the local community. Coastal communities that have beautiful areas will compete to provide tourist attractions in various ways, both through artificial and natural objects. With their unique characteristics, tourist attractions can attract the interest of visitors. Even though it is synonymous with pleasure, tourism activities on the coast also have risks. The thing that is possible to happen is the natural disaster of the tsunami. Tsunamis can cause severe damage to exposed areas and possibly cause physical disability to death for affected people. The increasing activity of community activities in the coastal environment has made it urgent to prevent victims of tsunami natural disasters [2].

In accordance with [3], for shorter return periods, such as 100 years, the primary tsunami threat is notably concentrated along the western coastlines of Nias, Mentawai, and Sumatra, as well as the southern coast of Java. Notably, a diminished relative hazard is observed along the mainland coast of Sumatra when islands, like Nias and Mentawai, are positioned offshore. Conversely, over longer return periods, the tsunami hazard in Eastern Indonesia, covering regions like North Papua and North Sulawesi, closely parallels the risk observed along the Sunda Arc. Based on various tsunami activities from 1922 to 2022 in Indonesia, it is recommended to develop research for multidisciplinary research in the future.

Density-Based Spatial Clustering of Applications with Noise, stands out as an effective method for discerning clusters within a dataset based on their density. Its effectiveness is particularly notable in its ability to identify clusters with irregular shapes, showcasing an advantage over certain other clustering techniques [4]. Unlike methodologies like k-means, which make assumptions about spherical clusters, DBSCAN excels at identifying clusters of arbitrary shapes. This adaptability proves invaluable when working with real-world datasets that may not adhere to predefined shapes. DBSCAN's versatility extends to various types of data, including spatial data, making it a popular choice in Geographic Information Systems (GIS), image analysis, and other fields that go beyond the conventional applications of clustering [5]. Its adaptability, automatic cluster determination, and robustness to noise contribute to DBSCAN's prominence as a powerful tool in the analysis and interpretation of complex datasets across diverse domains. Clustering tsunami characteristics using the Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm can be a supporting method in the review of tsunami hazard assessment by considering the history and geography of a particular area [6].

2. Material and Method

2.1. Data of Research

The dataset used is tsunami data in Indonesia, sourced from the National Centers For Environmental Information (NCEI) [7]. The research data consists of 135 tsunami events with 4 variables. The data period used in the restriction process is from 1822 to 2022. The dataset variables are in table 1.

Table 1. Variable Description.

Variable	Unit	Type
Focal Depth	Kilometer	Numeric
Earthquake Magnitude	Skala Richter	Numeric
Latitude	Degrees	Numeric
Longitude	Degres	Numeric

2.2. Clustering

It is a method used in data processing or data mining. Clustering aims to group data into a set based on data similarity. In the process, this unsupervised learning method will group data with the same characteristics into the same cluster [8]. Clustering algorithms that are currently available can be broadly divided into two types, namely partitioning and hierarchical. The partition clustering approach attempts to identify k partitions that optimize the function of specific criteria, with the squared error criterion being the most commonly employed [9].

$$E = \sum_{i=1}^{k} \cdot \sum_{p \in C_i} ||p - m||^2 \tag{1}$$

The squared error is a useful metric for assessing the degree of variance within the cluster across all partitions. The primary aim is to identify partition L that minimizes the squared error. Therefore, squared error clustering is focused on grouping k clusters together and keeping them as far apart as possible, which works well when the clusters are compact clouds that are relatively separated from each other. However, when there is a significant difference in the size or shape of the various clusters, as shown in Figure 1, the squared error method can cause large clusters to be split to minimize the squared errors. The squared error for (a) is higher for the three distinct clusters than for (b). The reduced squared error for (b) is due to the large cluster's separation being weighted by the many data points within it [10].

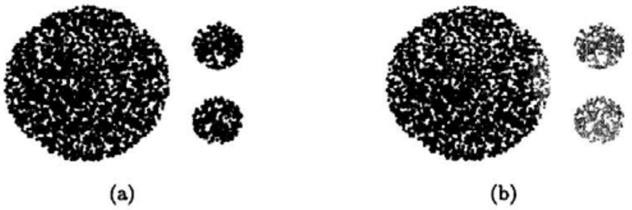

Fig 1. Large cluster division by partitioning algorithm [10].

2.3. *DBSCAN Algorithm*

In the Density-Based Spatial Clustering of Application with Noise (DBSCAN) algorithm, the process is clustering data based on density. This algorithm will use the minimum amount of data in a cluster with a minimum radius (ε) between data. If these elements are met, then the data can be grouped in one cluster [11].

The clustering process starts with several points in a data set (p ∈ D). The DBSCAN algorithm requires an approximate value of the density between data spaces. This algorithm will estimate the density value of each point in the dataset with the concept of neighborhood [12]. Where the value is,

$$N_{\epsilon}(p) = \{q\epsilon D \mid d(p,\ q) < \epsilon\} \tag{2}$$

where d is a calculated distance between two points. This density value will determine the types of data according to the requirements. As seen in Figure 2, if the Np density value is high or more than the minimum number of points (minPts), then that point can be classified as a core point [12].

$$core\,point\ = \{\ \mid N_{\epsilon}(p)\mid\ \geq\ minPts\ \} \tag{3}$$

Apart from core points, there are also types of border points. If a point being measured is not a core point but is still within the minimum range of the neighboring core point radius, it is a border point. If the point whose density is measured is not a core point or a border point, then the type is called a noise point [12].

$$border\,point\ = \{\ \mid N_{\epsilon}(p)\mid < minPts,\ \ q\epsilon \mid N_{\epsilon}\,(p)\mid\ \} \tag{4}$$

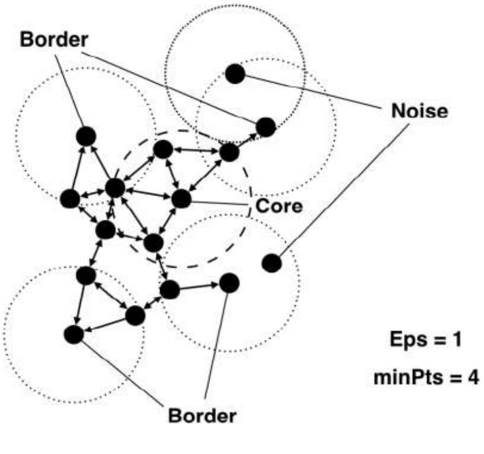

Fig 2. The concept of density in the DBSCAN algorithm [12].

Figure 2 explains how a point can be classified as a border point. Border points are formed when these points meet density-reachable with core points [13].

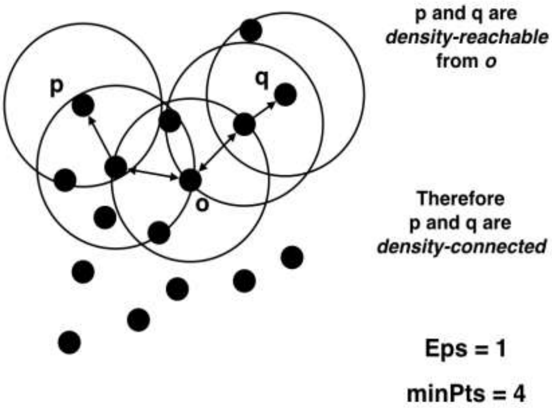
p
q
o
p and q are *density-reachable* from *o*
Therefore p and q are *density-connected*
Eps = 1
minPts = 4

Fig 3. Density connectivity concept illustration [12]

2.4. *Optics Algorithm*

The OPTICS algorithm is useful for datasets where the cluster structure cannot be accurately described using a single global density parameter [14]. This is often the case for real-world datasets, where different regions of the data space require varying local densities to reveal distinct clusters. For instance, in Figure 4, it is impossible to identify clusters A, B, C1, C2, and C3 simultaneously using a single global density parameter. Instead, the global decomposition of density parameters will only include clusters A, B, and C, or C1, C2, and C3, resulting in objects from clusters A and B being considered as noisy [14].

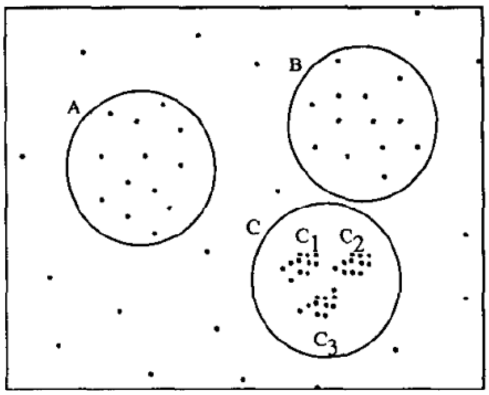

Fig 4. Distribution of data with different density parameters [15]

2.5. *Reachability Distance*

The reason behind utilizing the Reachability distance is not limited to measuring the distance between point p and its neighbors o. It is mostly used to account for statistical fluctuations in the distances of objects that are close to each other in the data space, such as in clusters or noise. Usually, the reachability distance is used to measure the distance to the core of each object, since it is a more accurate representation [16]. The MinPts parameter determines the strength of this effect. Higher values for MinPts mean that objects within the same area of the data space have more similar ranges. The OPTICS algorithm produces affordability plots with a similar smoothing effect, but it also reduces the impact of the so-called "single-link effect" on clustering [17].

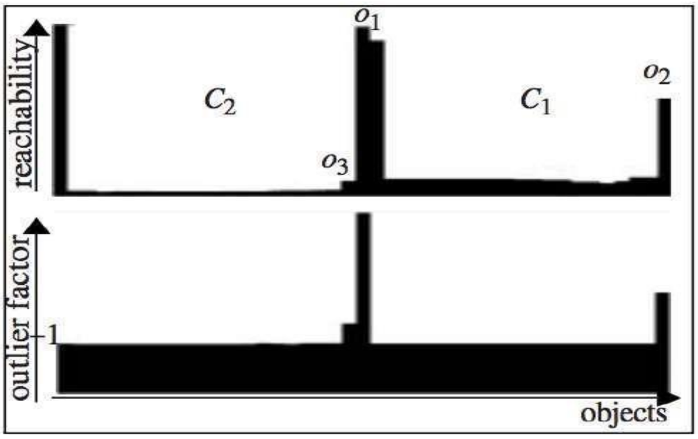

Fig 5. Reachability plots and factor outliners [17].

2.6. *Outliner*

Figure 6 illustrates that an object o in dataset D is considered an outlier DB (p,d) if a fraction of the objects in D are situated further than distance d from o. However, a disadvantage of this approach is that it takes a holistic view of the dataset. In reality, many datasets have more intricate structures where objects may only be outliers in relation to their local neighbors, while the overall distribution is disregarded [18]. Figure 6 offers a definition to capture such types of outliers.

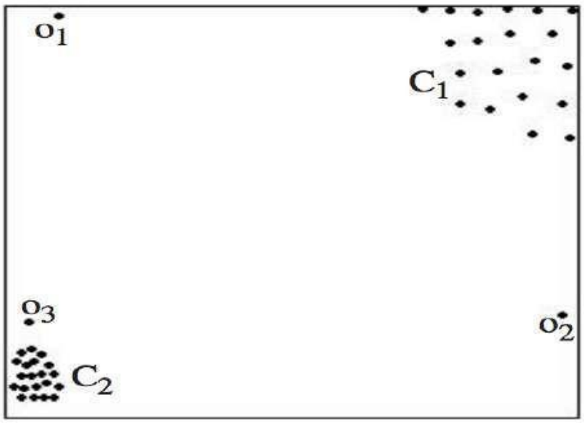

Fig 6. Outliner on a collection of datasets [18].

For example, suppose there is a data set that contains outliner objects. Figure 6 shows a data set with a total of 43 objects, which includes two clusters labeled C1 and C2, each containing 20 objects. In addition to these clusters, there are three other objects named o1, o2, and o3. According to the definition and intuition, o1, o2, and o3 are considered outliers, while the points that belong to clusters C1 and C2 are not outliers [18].

2.7. *Silhouette Score*

The silhouette value function in sci-kit-learn is to calculate the average silhouette coefficient of all samples. This value is obtained through the average intra-cluster distance (a) and the average distance of the nearest cluster (b) for each data. The silhouette coefficient is formulated by [19] :

$$\frac{b-a}{(a,b)} \tag{5}$$

The parameters that can be measured through silhouette values are:

- The silhouette value is close to +1, so the data points are in the correct cluster. The conclusion is that the quality of cluster modeling is very good.
- The silhouette value is close to 0, so several data points should belong to other clusters.
- The silhouette value is close to -1, so the data point is in the wrong cluster. The quality of the cluster models made is very poor. It is recommended to change the cluster-forming parameters.

3. Results and Discussion

3.1. Value of Radius (ε) *and MinPts*

Determination of the value of the minimum distance from each point to the core point is determined using the k-Nearest Neighbor algorithm with k = 2.

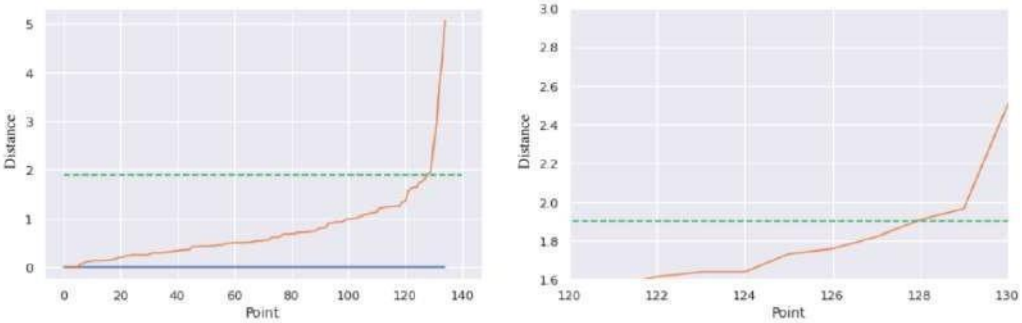

Figure 7. KNN in determining the value of radius (ε).

Based on Figure 7, it can be seen that the tipping point is at 1.9, which determines the value for further analysis. Meanwhile, determining the MinPts value can be done by taking into account the resulting Silhouette Coefficient value. In table 2 below, several Silhouette Coefficient values are displayed from the experimental variations.

Table 2. Variation of radius value with MinPts.

ε	MinPts	Silhouette	Cluster
1,9	8	0,46580	4
1,9	8	0,45430	5
1,9	7	0,47643	6
1,9	5	0,45339	7
1,9	4	0,49069	8
1,9	3	0,52927	9
1,9	3	0,50856	10
1,9	2	0,49557	11

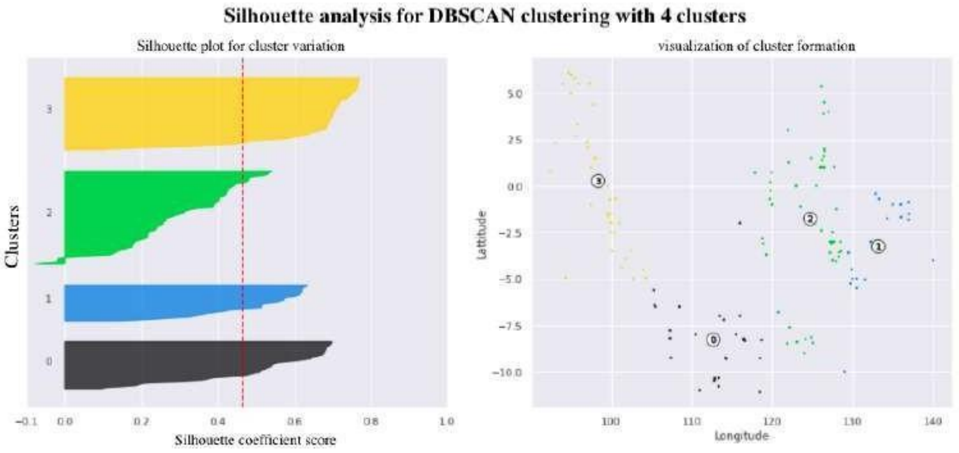

Figure 8. The analysis using silhouette method resulted in the formation of 4 clusters.

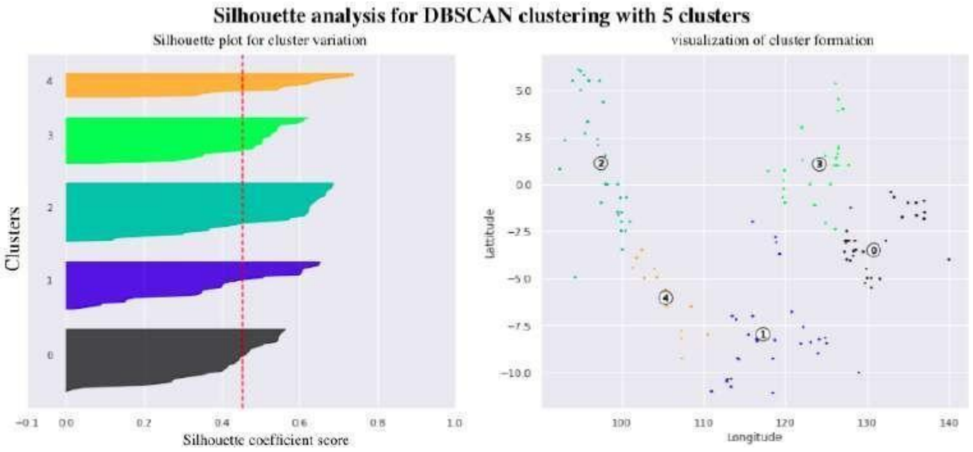

Figure 9. The analysis using silhouette method resulted in the formation of 5 clusters.

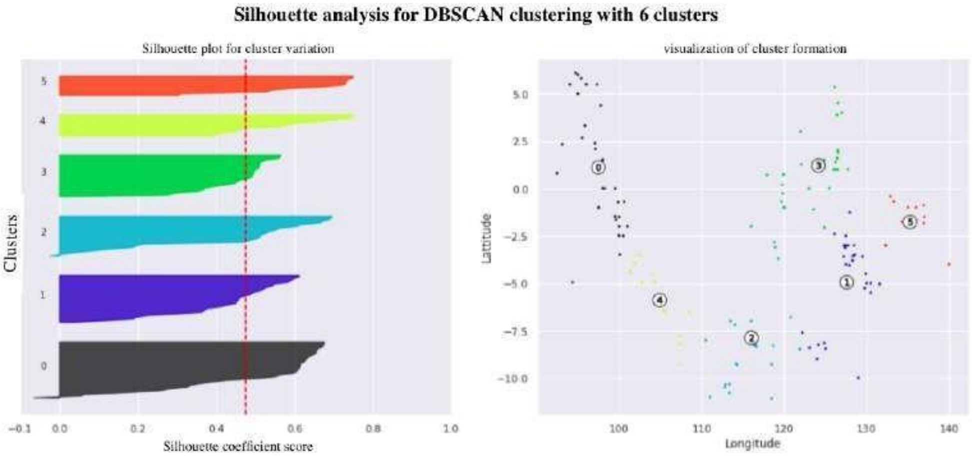

Figure 10. The analysis using silhouette method resulted in the formation of 6 clusters.

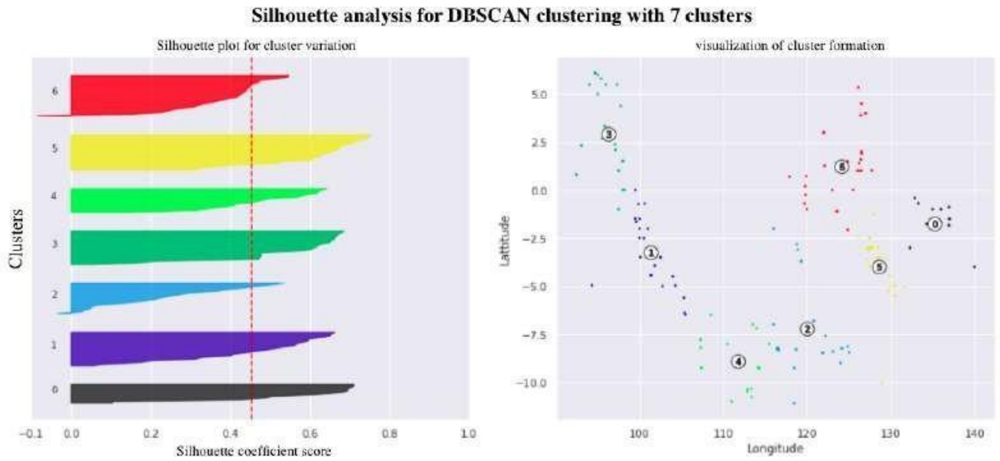

Figure 11. The analysis using silhouette method resulted in the formation of 7 clusters.

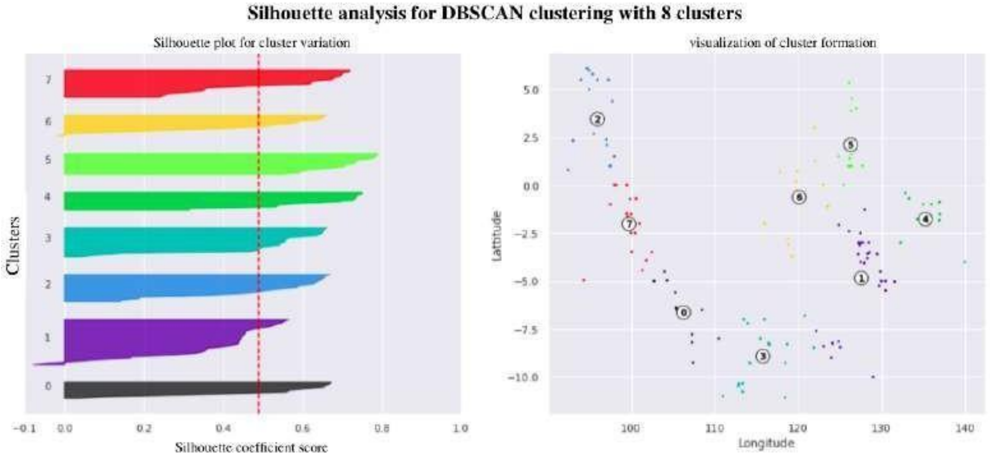

Figure 12. The analysis using silhouette method resulted in the formation of 8 clusters.

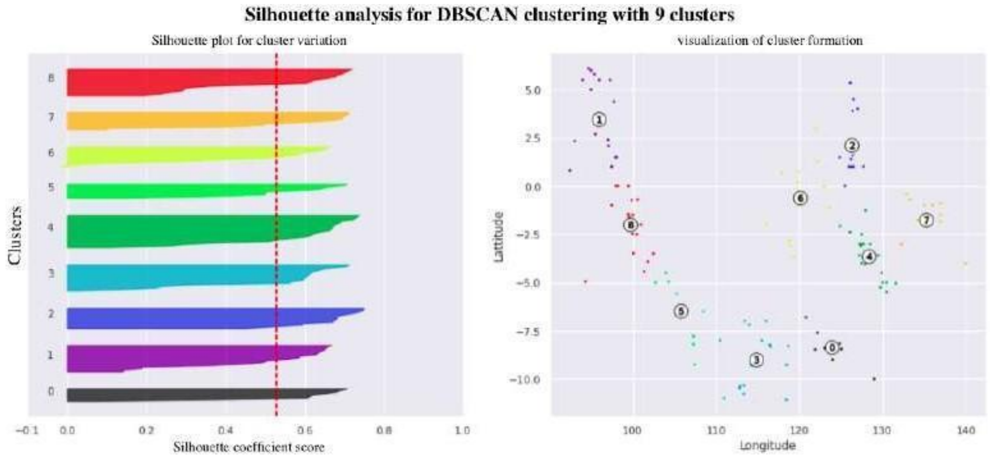

Figure 13. The analysis using silhouette method resulted in the formation of 9 clusters.

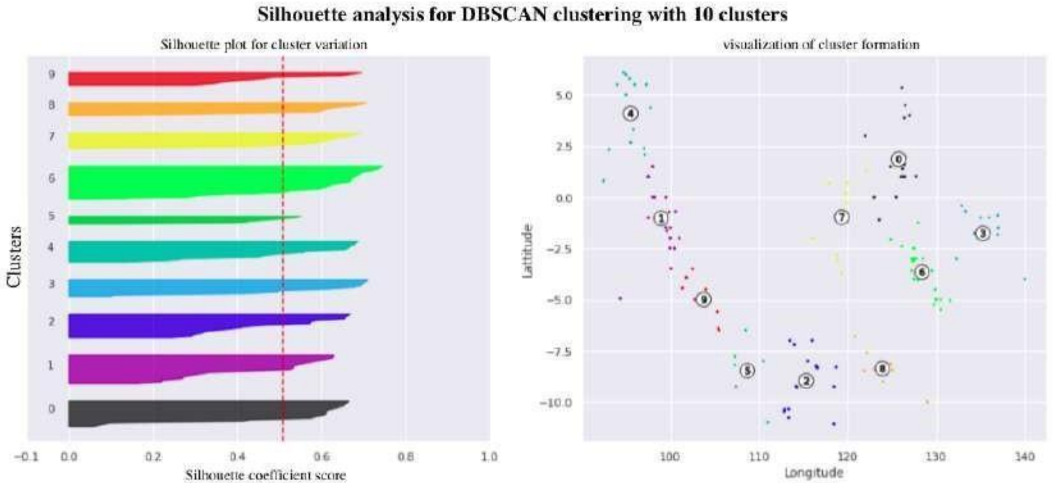

Figure 14. The analysis using silhouette method resulted in the formation of 10 clusters.

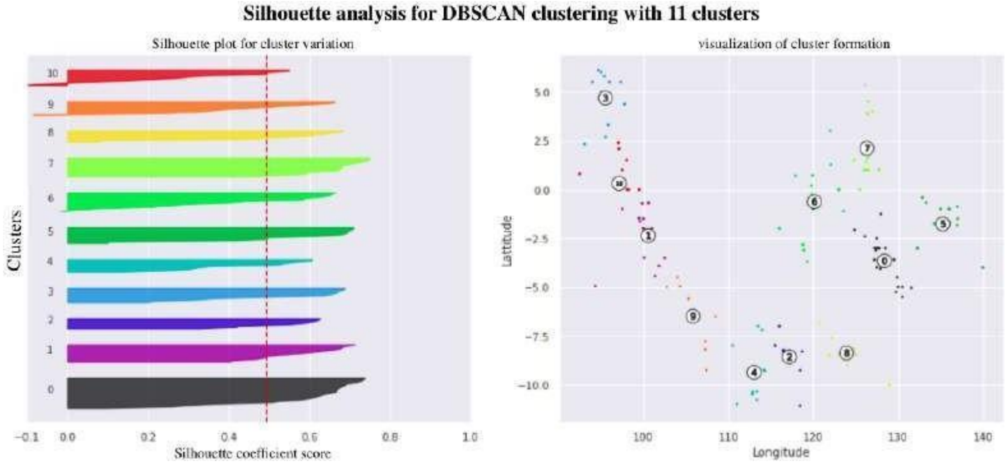

Figure 15. The analysis using silhouette method resulted in the formation of 11 clusters.

The addition of the MinPts value will determine the number of groups formed. The higher the MinPts value, the better the cluster quality will be. This quality can be determined through the Silhouette Coefficient. However, there are positions where cluster quality will deteriorate as the MinPts value increases. As shown in table 2, if MinPts is greater than 3, the Silhouette value will decrease. This indicates that there will be a decrease in cluster quality at the MinPts value. Meanwhile, in taking the MinPts value, it is determined through the largest Silhouette value.

3.2. Clustering with DBSCAN

These values (ε) and MinPts (ε =1.9 and MinPts = 3) will later be used as input in running the clustering algorithm on DBSCAN. In the process, 10 clusters were formed from 135 tsunami event data. Each number of tsunami event data in each cluster is shown in table 4 and their distribution is in Figure 16.

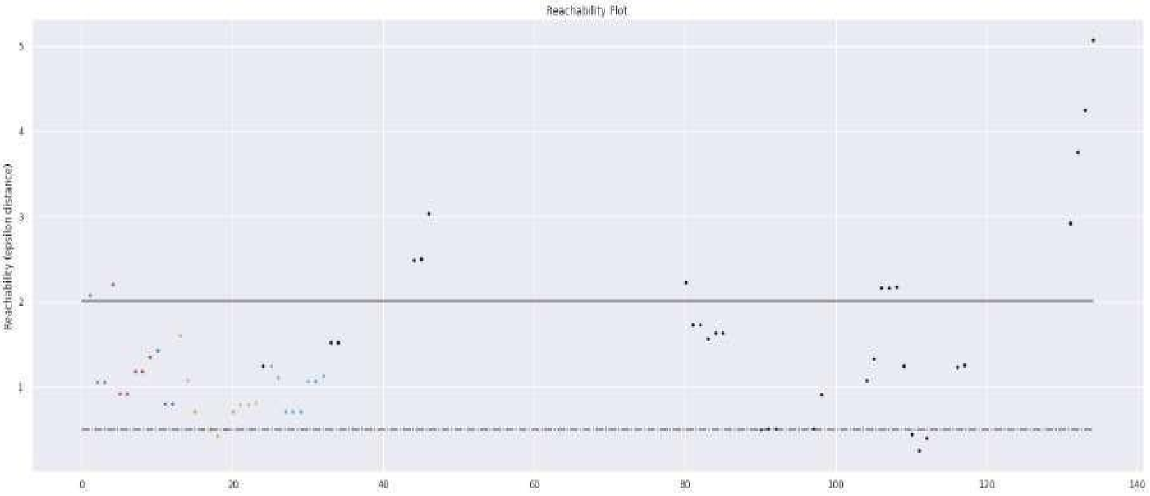

Figure 16. Reachability distance plots.

Table 3. DBSCAN model output with MinPts=3 and ε=1.9

Output	Cluster	Amount of data
0	Cluster 1	39
1	Cluster 2	26
2	Cluster 3	13
3	Cluster 4	14
4	Cluster 5	9
5	Cluster 6	8
6	Cluster 7	5
7	Cluster 8	4
8	Cluster 9	3
9	Cluster 10	3
-1	Outliners	11

3.3. Clustering with OPTICS

The value and MinPts used in this algorithm are the same as the previous algorithm. Table 4 is the result of clustering with OPTICS and figure 17 is the distribution of the clusters formed. In addition, table 6 also presents the amount of data in each cluster on OPTICS clustering.

Table 4. Parameter input and clustering results with OPTICS.

Parameter	
MinPts	3
ε	1,9
Output	
Cluster	18
Outliner	32

Table 5. The output of OPTICS clustering results

Output	**Cluster**	**Amount of data**
-1	Outliners	32
0	Cluster 1	4
1	Cluster 2	6
2	Cluster 3	3
3	Cluster 4	11
4	Cluster 5	8
5	Cluster 6	9
6	Cluster 7	6
7	Cluster 8	7
8	Cluster 9	3
9	Cluster 10	8
10	Cluster 11	3
11	Cluster 12	6
12	Cluster 13	4
13	Cluster 14	4

14	Cluster 15	5
15	Cluster 16	3
16	Cluster 17	4
17	Cluster 18	9

3.4. Cluster Validation

The study uses the silhouette metric to evaluate the internal quality of the clustering models. The evaluation results are reported in Table 6, which indicates that the DBSCAN algorithm outperforms the OPTICS algorithm. This can be observed from the higher silhouette scores and the fewer number of outliers in the DBSCAN model. A higher silhouette score suggests a better quality of the clustering model. These results can be used to compare with other parameters in future research.

Table 6. DBSCAN comparison with OPTICS.

	DBSCAN	**OPTICS**
Amount of cluster	9	18
Outliners	11	32
Koef. Silhouette	0,378835	0,242401

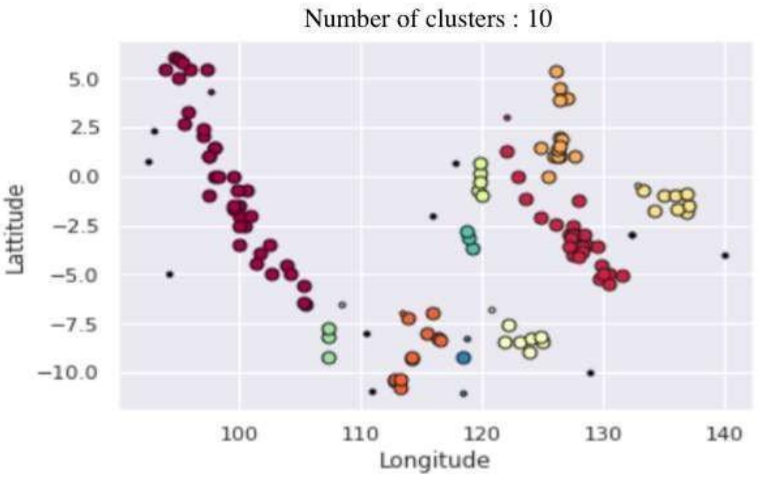

Figure 17. Clustering plots with DBSCAN.

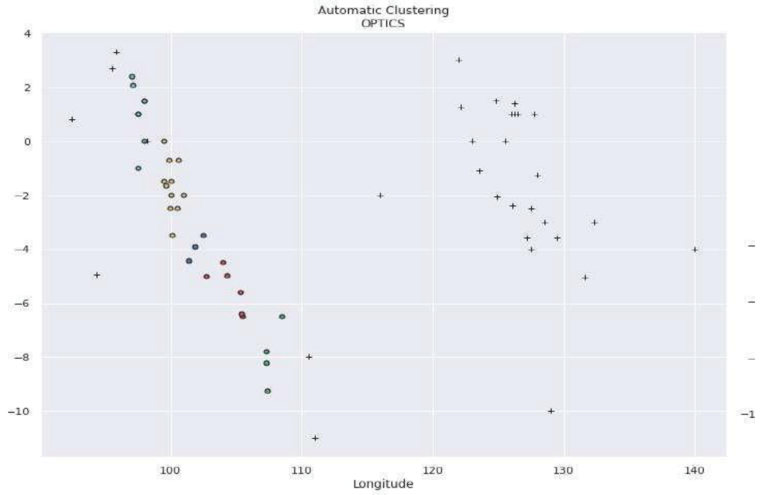

Figure 18. Clustering plots with OPTIC.

3.5. *Tsunami Characteristic Clusters*

Clustering the characteristics of the tsunami based on the variables in section 3.1 resulted in three clusters. With the KNN algorithm, the value of eps = 10 is obtained. The best silhouette value calculation is found in cluster 3 with a value of 0.6341.

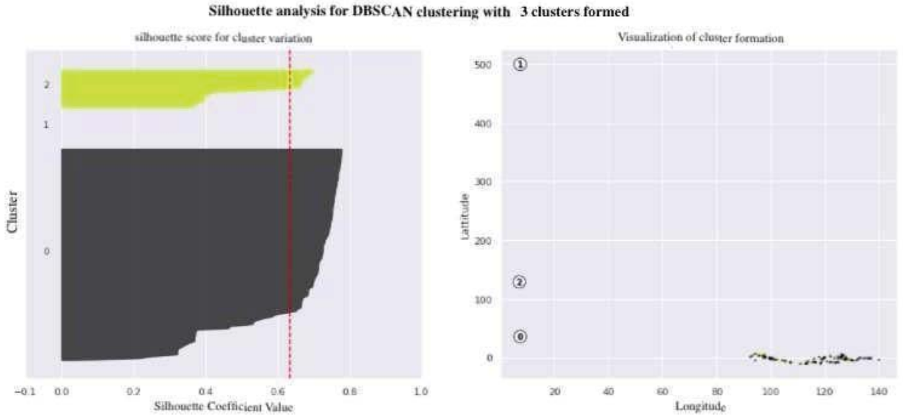

Figure 19. Calculation of eps values.

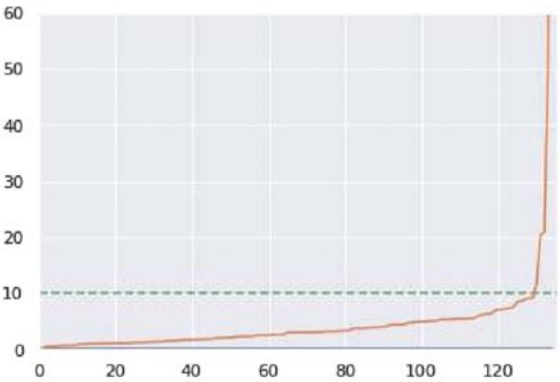

Figure 20. Silhouette calculation results in.

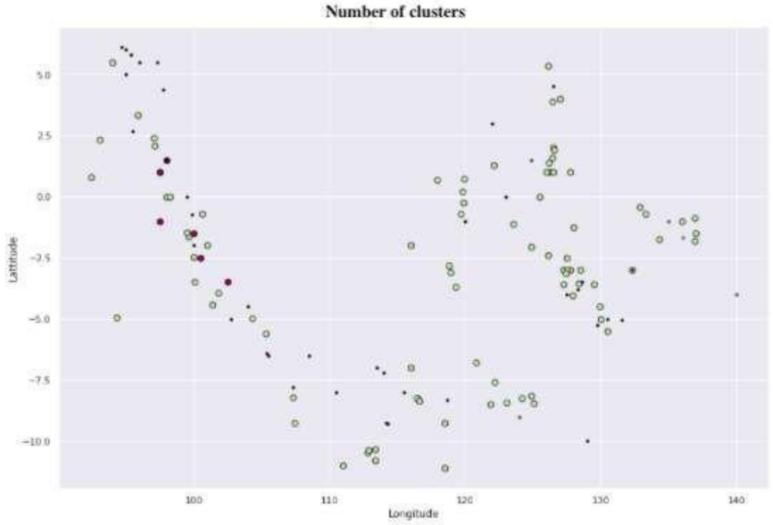

Figure 21. Results of clustering characteristics of the tsunami.

4. **Conclusion**

The study concludes that the DBSCAN algorithm is superior to the OPTICS algorithm. Using the DBSCAN algorithm with ε = 1.9 and MinPts = 3, 10 clusters of tsunami-prone areas and 3 clusters of tsunami characteristics in Indonesia were identified, with a silhouette score of 0.378835. The clustering results identified several areas with potential for tsunamis, including the coasts of Sumatra Island, the Sunda Strait, the Java Sea, the Bali Strait, the Nusa Tenggara Sea, the Sulawesi Sea, the Maluku Sea, and the West Papua Sea.

References

[1] Satria, A., “Introduction to Sociology of Coastal Communities”. 2015. Jakarta : Yayasan Pustaka Obor Indonesia.

[2] Pamungkas, M.Si., D.N., Rush Listiyani, S.E., Ak., M.Si, & Saptono, M.M., D.R.H., “Improvement of Coastal Tourism in Java Island”. 2019. Nugra Media.

[3] Horspool, N., Pranantyo, I., Griffing, J., & Latief, H., “A probabilistic tsunami hazard assessment for Indonesia”. Nat. Hazard Earth System Science. 2014.

[4] Ashour, W., & Sunoallah, S., “Multi Density DBSCAN. In: Yin, H., Wang, W., Rayward-Smith”, Intelligent Data Engineering and Automated Learning Series V, vol 6936. Springer, Berlin, Heidelberg. 2011.

[5] Augsondit, T., & Khemniwat, T., & Sathirasattayanon, P., & Kaewcharuay, P., & Galajit, K., & Karnjana, J., & Usanavasin, S., “Rockfall Isolation Technique Based on DC-DBSCAN with k-Means Clustering and k-Nearest Neighbors Algorithm”, 2023.

[6] Arafat, I., & Hariyadi, M., & Santoso, I., & Crysdian, C., “Earthquake Clustering in Regional VII Using DBSCAN Approach”, Journal of Information Technology and Computer Science. 2023.

[7] National Geophysical Data Center/World Data Service: NCEI/WDS Global Historical

Tsunami Database. NOAA National Centers for Environmental Information. doi:10.7289/V5PN93H7.

[8] Mashat, A.F., & Fouad, M. M., & Yu, P. S., & Gharib, T. F., “Efficient Clustering Technique for University Admission Data,” International Journal of Computer Applications, vol. 45, no. 23, pp. 39–42, 2012.

[9] Nafuri, A. F. M., & Sani, N. S. & Zainudin, N. F. A. & Rahman, A. H. A., & Aliff, M., "Clustering Analysis for Classifying Student Academic Performance in Higher Education”, J.Applied Science, vol. 12, no. 19, pp. 1-22, 2022, doi: 10.3390/app12199467.

[10] Guha, S., Rastogi, R., & Shim, K., “CURE: An Efficient Clustering Algorithm for Large Databases”. 2019.

[11] Deng, D., “DBSCAN Clustering Algorithm Based on Density”, 7th International Forum on Electrical Engineering and Automation (IFEEA), pp. 949–953, 2020. doi: 10.1109/IFEEA51475.2020.00199.

[12] Hahsler, M., Piekenbrock, M., & Doran., “Fast Density-Based Clustering With R”. Journal of Statistical Software. 2019. vol.91

[13] Khan, K., Rehman, S. U., Fong, S., & S, S., “DBSCAN: Past, Present and Future”, 2014.

[14] Tang, Chunhua & Wang, Han & Wang, Zhiwen & Zeng, Xiangkun & Yan, Huaran & Xiao, Yingjie.,“An improved OPTICS clustering algorithm for discovering clusters with uneven densities”, Intelligent Data Analysis, vol. 25, pp.1453-1471,2021. doi:10.3233/IDA-205497.

[15] Ankerst, M., Breuning, M. M., Kriegel, H. P., & Sander, J., “OPTICS: Ordering Points To Identify the Clustering Structure”. 2019.

[16] Grigoli, Francesco & Scarabello, Luca & Boese, Maren & Wiemer, Stefan & Clinton, John., “How the DBSCAN Algorithm works: A simple introduction using sketches”, 2018.

[17] Sibson, R., “SLINK: an optimally efficient algorithm for the single-link cluster method”. The Computer Journal. Vol. 16. No.1. pp.30-34.

[18] Breuning, M. M., Kriegel, H. P., Raymond, T. N., & Sander, J., “OPTICS-OF: Identifying Local Outliers”. 2019

[19] Shahapure, K. R., & Nicholas, C., “Cluster Quality Analysis Using Silhouette Score”, 2020.