Mathematical Journal of Modelling and Forecasting
Vol. 3, No. 1, June 2025, pp. 8-15

Earthquake Point Clustering in Sumatra Island using Spatio-Temporal Density-Based Spatial Clustering Application with Noise (ST-DBSCAN) Algorithm

Muthiara Hazimah Putri[1,*], Devni Prima Sari[2]

[1,2] Departement of Mathematics, Universitas Negeri Padang, Indonesia
[*] muthiaraputri04@gmail.com

Abstract. Earthquakes are one of the natural disasters that often occur in Indonesia, especially on the island of Sumatra. Earthquakes become a frightening spectre because they cannot be predicted when they will come, where they will be located, and how strong the vibrations are, so they often cause damage and casualties. To minimise losses due to earthquakes, it is necessary to divide areas easily affected by earthquakes. One method that can be used to divide these areas is clustering techniques. This study uses a clustering method, namely Spatio Temporal-Density Based Spatial Clustering Application with Noise (ST-DBSCAN), on the dataset of earthquake points on the island of Sumatra in 1917-2023. This method uses a spatial distance parameter (ε_1= 0.28), temporal distance parameter (ε_2= 180), and minimum number of cluster members (MinPts = 7) with a silhouette coefficient of 0.0991, resulting in 145 clusters with 15 large clusters and 4922 noises. The epicentres are primarily located in Siberut Island, Tanah Bala Island and its surroundings, the Indian Sea opposite Nias Island, the Sea around the Mentawai Islands, Enggano Island and its environs, Simaulue Regency, and Enggano Island and the Sea around it. The most common type of spatio-temporal pattern found is the occasional pattern type.

Keywords: Clustering, Data Mining, Earthquake, ST-DBSCAN

1 Introduction

Earthquakes are vibrations or shocks that occur on the Earth's surface due to the sudden release of energy from within the Earth. Earthquakes are natural disasters caused by a sudden release of energy in the Earth's crust, accompanied by vibrations or shocks on Earth's surface, causing seismic waves. Earthquakes are a commoditised mechanism of large-scale movement of the Earth's plates as a result of plate collisions that cause extreme swings and vibrations that produce seismic energy [1]. Shifts in plate tectonics can cause pressure and stress on rocks beneath the Earth's surface. When this pressure and stress exceed the strength of the rock, the rock will break and collapse [2]. Earthquakes are a serious problem in Indonesia because they cannot be predicted when, where, or how strong they will be.

Sumatra is one of the islands located along the Indo-Australian and Eurasian plates' meeting points. A 1900 km-long fault zone passes through Sumatra, near or within an active volcanic arc. Sumatra is split from Semangko Bay to Banda Aceh by active faults, which have high seismic activity, making the island vulnerable to earthquakes that cause great danger to densely populated areas near active fault lines [3]. In recent years, earthquakes have occurred in Indonesia, causing significant impacts, both material and in terms of casualties. Information that explains natural disasters through a data mining approach is needed because of the many natural disasters that occur on the island of Sumatra [4].

Earthquakes can cause very severe and dangerous damage, especially in earthquake-prone areas. A fundamental thing that can be done is to divide the areas that most frequently experience earthquakes to minimise the risk of casualties and infrastructure damage. Clustering can be done to facilitate the division of earthquake areas. Clustering techniques include data mining techniques that group data based on similarities in data characteristics [5]. Clustering aims to group objects with similarities in the same cluster. The higher the similarity within a group and the greater the difference between groups, the better or different the grouping.

One method that can be used is Spatio Temporal-Density Based Spatial Clustering Application with Noise (ST-DBSCAN). The ST-DBSCAN algorithm is one of the clustering methods that can process temporal data using distance in both spatial and temporal aspects. This allows it to effectively capture patterns that evolve, such as object movement or changes in sensor readings. The ST-DBSCAN algorithm is nonparametric in unsupervised learning, so it does not require assumptions in its work and is one of the best methods for determining clusters from large spatial databases [6]. The ST-DBSCAN algorithm is one of the clustering methods that can process temporal data using distance in both spatial and temporal aspects. This allows it to effectively capture patterns that evolve, such as object movement or changes in sensor readings.
Several researchers have researched clustering using the ST-DBSCAN algorithm. Research with the ST-DBSCAN algorithm in other natural disaster fields[7], [8], [9], [10]. Research with the ST-DBSCAN algorithm in the health sector [11], [12]. As well as research with the ST-DBSCAN algorithm in the field of human indices [13].

2 Research Methods

This research uses earthquake point data on the island of Sumatra published from the United States Geological Survey (USGS) website, with 11,049 earthquake points. The USGS publication's data includes tables describing earthquake attributes, totalling 22 attributes such as time, longitude, latitude, mag, magtype, nst, gap, and others. This study uses three attributes: longitude, latitude, and time.

The data analysis techniques performed in this study are as follows:

1) Processing data.
    a. Change the format of the data column.
    b. Data reduction.
2) Determining the parameter values ε and MinPts.
3) Calculating the Euclidean distance between objects based on spatial aspects (longitude and latitude) and based on temporal aspects (date of earthquake occurrence).
    For the spatial aspect of the Euclidean distance equation [14]:

    $$d(i,j) = \sqrt{\left(\left(X_{longi} - X_{longj}\right)^2 + \left(X_{lati} - X_{latj}\right)^2\right)} \quad (1)$$

    where:
    long i = *longitude* data to$-i, i = 1,,..,n$
    lat j = *latitude* data to$-j, j = 1,,..,n$

    For the temporal aspect, the Euclidean distance equation is modified into:

    $$d_t\,(i,j) = \left|X_{date_i} - X_{date\,j}\right| \quad (2)$$

    where:
    X_{date} is the object in the date column and is a one-dimensional object.
4) Performing cluster analysis with the ST-DBSCAN algorithm.
5) Measuring cluster quality using the silhouette coefficient [15].

    $$s(i) = \frac{b(i)-a(1)}{\max\{a(i),b(i)\}} \quad (3)$$

    where:
    $s(i)$ = The silhouette coefficient value of i.
    $a(i)$ = Average distance of object i to all other objects in the cluster a.
    $b(i)$ = The average minimum distance from object i to all other clusters (that are not clusters of object (i))
6) Analysing spatio-temporal patterns based on the best parameters of ε and MinPts.
7) Visualising and interpreting the results of the clusters formed from the ST-DBSCAN method.

3 Results and Discussion

3.1 Preprocessing Data

a. Formatting the date column.

This formatting is done so that when additional data is present before the first date, the data used at this time does not cause the date variable to be -1. The date format changes can be seen in Table 1.

Table 1. Transform Date Format to Number.

No.	Date (GMT)	Date (number)
1	2023-12-23 (10:26:19)	45283
2	2023-12-21 (23:10:44)	45282
3	2023-12-16 (06:45:55)	45276
4	2023-12-13 (11:08:55)	45273
5	2023-12-11 (10:16:39)	45271
⋮	⋮	⋮
11.046	1918-02-12 (03:00:34)	6618
11.047	1917-11-04 (12:03:34)	6519
11.048	1917-04-16 (18:45:11)	6317

b. Data Reduction

At this stage, the number of variables to be used is reduced. The initial 22 variables included longitude, latitude, date of the earthquake, earthquake strength, depth of the earthquake, etc. At this stage, the data to be used was reduced to longitude, latitude, and date of the earthquake.

3.2 Determination of ε and MinPts Parameters

Determination of the optimal ε_1 value, selected based on the k-dist graph using KNN. This research's k values are k=3, k=4, k=5, and k=7. In Figure 1, Figure (a) is a k = 3 graph, Figure (b) is a k = 5 graph, and Figure (c) is a k = 7 graph is a k-dist graph for determining the parameter ε_1. The value of ε_1 obtained on the line with significant curvature. If using Minpts $= 3$, the optimal ε_1 is 0.18 to 0.28; if using MinPts $= 5$, the optimal ε_1 is 0.22 to 0.32; if using MinPts $= 7$, the optimal ε_1 is 0.26 to 0.35. The ε_1 values to be used are the upper, lower, and center values of the k-dist graph. While the selected ε_2 values are 7 days, 30 days, and 180. The value of ε_1 selected in this study can be seen in Table 2.

Table 2. The Value of ε_1

MinPts	ε_1
	0.18
3	0.23
	0.28
	0.22
5	0.27
	0.32
	0.26
7	0.31
	0.36

3.3 Euclidean Distance

In the cluster formation stage, the distance between objects must be calculated. The calculation uses the Euclidean distance method, which is based on spatial aspects (longitude and latitude) and temporal aspects (date of the earthquake), as shown in Table 3. Table 3 shows that the distance between all observations is calculated based on spatial aspects (longitude and latitude) and temporal aspects (date of the earthquake). For example, the second row and first column show that the distance between objects is approximately 3.334. This means the distance between the first and second objects is about 3.334.

Table 3. Euclidean Distance Calculation

(i,j)	*d(i,j)* **Spatial**	*d(i,j)* **Temporal**
(1,1)	0	0
(1,2)	3,334	1
(1,3)	0,437	7
(1,4)	3,583	10
(1,5)	2,249	12
⋮	⋮	⋮
(11.047,11048)	2,467	100
(11.048,11049)	5,459	202
(11.049,11049)	0	0

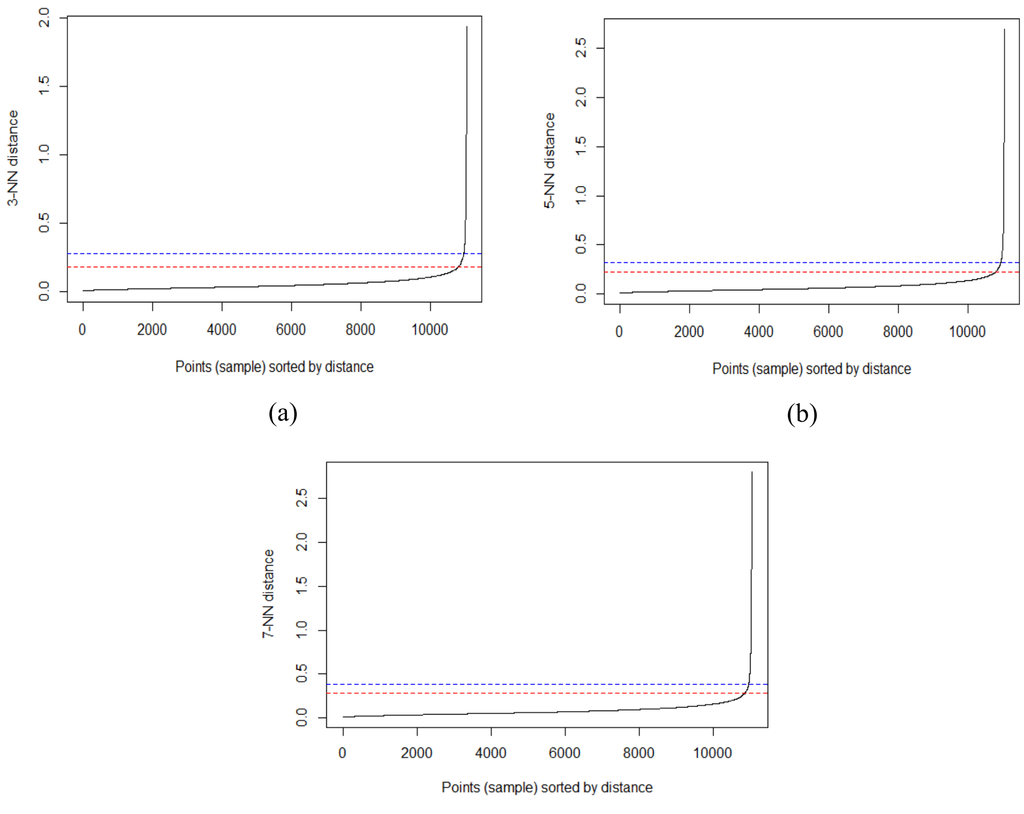

Fig 1. Parameter Determination of ε_1 against Graph (a) $k = 3$, (b) $k = 5$, and (c) $k = 7$

3.4 Cluster Formation

After obtaining the pairs of parameters ε_1, ε_2, and MinPts on the earthquake point data on the island of Sumatra, then process the data using the ST-DBSCAN algorithm using R software. After obtaining the cluster results, the parameter with the highest silhouette coefficient value for each MinPts is selected in Table 4. The best parameter selection using the ST-DBSCAN algorithm is the silhouette coefficient value obtained between 0 and 1. In Table 5, it can be seen that the best parameters using the ST-DBSCAN algorithm in this study are selected $\varepsilon_1 = 0.28$, $\varepsilon_2 = 180$ and MinPts $= 7$ gets the highest silhouette coefficient value of 0.0991 with the number of clusters 145 and noise 4922 points.

3.5 Analisis Pola Spatio-Temporal

According to [16], large clusters that can be analyzed for patterns are clusters with a minimum number of points of 180, while clusters with less than 180 points are small clusters whose patterns cannot be detected. Based on the best parameters, where $\varepsilon_1 = 0.28$, $\varepsilon_2 = 180$, and MinPts $= 7$, there are 15 large clusters that have more than 30 points.

Each large cluster will be divided into periods, because the best parameter in this study has a value of $\varepsilon_2 = 180$. The period is 180 days, and the location of the distribution of earthquake points in one period is compared with the next period. For example, cluster 113 obtained four periods, as shown in Figure 3.

In Figure 3, it can be seen that the earthquake points are also irregular, as in period 1, the earthquake points are pretty dense and centered in certain areas, indicating that earthquakes often occur in the exact location, period 2 to period 4 show the distribution of earthquake points rarely and randomly scattered, so it can be said that cluster 113 has an irregular pattern. The earthquake occurred around West Aceh Regency and the Aceh Sea in Banda Aceh Province. There are seven occasional patterns, and each has its cluster centre. Siberut Island, Tanah Bala Island and its surroundings, the Indian Sea across Nias Island, the Sea around Mentawai Islands, Enggano Island and its environs, Simaulue Regency, and Enggano Island and its surrounding Sea are the areas with the most cluster centres of earthquake points.

Table 4. Clustering Result on Earthquake Point Data Experiment

ε_1	ε_2	MinPts	Result		
			Cluster	Noise	Silhoutte coefficient
0.18	7	3	741	6552	-0.09406501
0.18	7	5	186	8075	-0.1420485
0.18	7	7	90	8650	-0.04153464
0.18	30	3	834	5370	-0.04872923
0.18	30	5	197	7056	-0.1249481
0.18	30	7	97	7661	-0.08319615
0.18	180	3	888	3198	-0.04557093
0.18	180	5	265	5218	-0.05784932
0.18	180	7	130	6089	-0.04716562
0.22	7	3	776	6206	-0.1500897
0.22	7	5	188	7801	-0.2002198
0.22	7	7	96	8347	-0.1057129
0.22	30	3	893	4930	-0.07529446
⋮	⋮	⋮	⋮	⋮	⋮
0.28	30	3	962	4393	-0.1086204
0.28	30	5	214	6396	-0.1175298
0.28	30	7	104	6996	-0.06262691
0.28	180	3	790	2021	-0.1796224
0.28	180	5	293	3896	-0.07933323
0.28	180	7	145	4922	0.09910172
0.31	7	3	840	5714	-0.175355
0.31	7	5	202	7367	-0.2108252
0.31	7	7	110	7893	-0.1722351
0.31	30	3	986	4191	-0.1211619
⋮	⋮	⋮	⋮	⋮	⋮
0.36	7	7	112	7745	-0.1882226
0.36	30	3	1023	3863	-0.1389615
0.36	30	5	235	6009	-0.1191624
0.36	30	7	119	6663	-0.08839106
0.36	180	3	647	1475	-0.2976715
0.36	180	5	246	3093	-0.1848476
0.36	180	7	156	4114	-0.07373127

Fig 2. Cluster 113 Earthquake Point Pattern.

Table 5. Large Cluster

Cluster	Number of Point
1	51
9	95
14	41
51	2798
72	229
89	123
97	603
108	51
113	43
114	59
115	325
119	48
121	37
122	471
142	40

4 Conclusion

The results of clustering using the ST-DBSCAN method show that the best optimal parameter values are $\varepsilon_1 = 0.28$, $\varepsilon_2 = 180$ and MinPts $= 7$ with a silhoutte coefficient of 0.0991, resulting in 145 clusters with 15 large clusters and 4922 noise. The epicentres were primarily located in Siberut Island, Tanah Bala Island and its surroundings, the Indian Ocean across Nias Island, the Sea around Mentawai Islands, Enggano Island and its surroundings, Simaulue Regency, and the sea around Enggano Island and its environs. The most common type of spatio-temporal pattern found is occasional.

References

[1] E. R. Hartuti, *Buku Pintar Gempa*. DIVA Press, 2009.

[2] A. Fauzi, *Fisika : Bencana Alam*. UNP Press, 2013.

[3] M. Affan, M. Syukri, L. Wahyuna, and H. Sofyan, “Spatial Statistic Analysis of Earthquakes in Aceh Province Year 1921-2014: Cluster Seismicity,” vol. 5, no. 2, 2016.

[4] A. F. Boer, M. A. Haris, and R. Wasono, “Pengelompokan Daerah Rawan Bencana di Pulau Sumatera dengan Metode Density Based Spatial Clustering of Applications with Noise (DBSCAN),” vol. 6, 2023.

[5] H. Cui, W. Wu, Z. Zhang, F. Han, and Z. Liu, “Clustering and application of grain temperature statistical parameters based on the DBSCAN algorithm,” 2021.

[6] M. N. Gaonkar and K. Sawant, “AutoEpsDBSCAN : DBSCAN with Eps Automatic for Large Datasets,” vol. 2, no. 2, 2013.

[7] A. Johar, A. Vatresia, and I. A. Donny, “Implementasi Metode Spatio-Temporal Clustering dengan Aalgoritma ST-DBSCAN Pada Titik Api Kebakaran Hutan Indonesia (2015-2020),” *Jurnal Rekursif*, vol. 11, no. 1, 2023.

[8] N. A. Putri, T. W. Utami, and R. Wasono, “Spatial Temporal Density-Based Spatial Clustering Applications With Noise (ST-DBSCAN) Untuk Pengelompokan Titik Kebakaran Hutan dan Lahan di Provinsi Riau Pada Tahun 2021,” vol. 6, 2023.

[9] A. N. Fahira and R. Nooraeni, “Optimasi Parameter ST-DBSCAN dengan KNN dan Algoritma Genetika Studi Kasus: Data Bencana Alam di Pulau Jawa 2021,” vol. 11, no. 1, 2023.

[10] G. R. W.S and H. Fahmi, “Implementasi Metode ST-DBSCAN untuk Pengelompokan Pola Persebaran Titik Api pada Data Kebakaran Hutan di Indonesia,” vol. 3, no. 5, 2024.

[11] K. B. Chimwayi and J. Anuradha, “Clustering West Nile Virus Spatio-temporal data using ST-DBSCAN,” *Procedia Computer Science*, vol. 132, pp. 1218–1227, 2018, doi: 10.1016/j.procs.2018.05.037.

[12] S. H. Hastuti, A. Septiani, Hendrayani, and W. P. Nurmayati, “Penerapan Metode OPTICS dan ST-DBSCAN untuk Klasterisasi Data Kesehatan,” vol. 2, no. 1, 2024.

[13] M. I. P. Eriansya and M. Syafrullah, “Implementasi Algoritma ST-DBSCAN dan K-MEANS Untuk Pengelompokan Indeks Pembangunan Manusia Kabupaten/Kota Pulau Jawa Tahun 2014-2016 Berbasis Web Di Badan Pusat Statistik,” vol. 1, no. 3, 2018.

[14] J. Han, M. Kamber, and J. Pei, *Data Mining : Concepts and Techniques*. 2012.

[15] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. 2006.

[16] C. Pöelitz and N. Andrienko, “Finding arbitrary-shaped clusters with related extents in space and time,” 2010.