

Optimizing seismic sequence clustering with rapid cube-based spatiotemporal approach

Silviya Hasana¹, Wina Permana Sari², Rojali³, Devi Fitrihanah³

¹Independent Researcher, Jakarta, Indonesia

²Department of Computer Science, Faculty of Computer Science, Bina Nusantara University, Jakarta, Indonesia

³Department of Computer Science, BINUS Graduate Program, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Nov 4, 2023

Revised Jul 5, 2024

Accepted Jul 26, 2024

Keywords:

Clustering

Density-based

Earthquake

Seismic sequence

Spatiotemporal

ABSTRACT

Due to their extensive volume and range of features, seismic data is regarded as highly complex data. Earthquakes that typically composed of foreshocks, mainshocks, and aftershocks, exhibit a unique sensitivity to temporal dimension, a characteristic that differs them from other natural hazards. Foreshocks and aftershocks that emanate from a similar epicenter, often display temporal patterns that contribute significantly to determining a sequence. This study introduces a density cube-based approach to cluster spatiotemporal seismic data. It addresses spatial irregularities observed in earthquake clusters and incorporates temporal aspects, acknowledging that seismic events originating from a similar epicenter could occur in separate time frames. We achieved the highest Silhouette score of 0.935 in daily-based clustering and 0.782 in weekly-based clustering. Notably, our analysis reveals a trend where weekly clustering lambda λ tend to be lower ($\lambda=0.01$) than in daily clustering ($\lambda=0.1$, $\lambda=0.5$), thus emphasizing the significance of temporal granularity where daily clustering requires higher λ to capture rapid fluctuations, while weekly clustering benefits from lower λ to cover broader trends. These findings enhance the understanding of the nuanced interplay of temporal dynamics in seismic sequence analysis.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Silviya Hasana

Independent Researcher

Jakarta, Indonesia

Email: silviyahasana@gmail.com

1. INTRODUCTION

Earthquakes are abrupt releases of energy in the earth's lithosphere with complex, nonlinear and unpredictable nature over high-dimensional space and time [1]. An earthquake sequence typically composed of foreshocks, mainshocks, and aftershocks. Foreshocks are minor earthquakes prior to the mainshock, while aftershocks follow the mainshock with a gradual decrease in frequency and amplitude. Although foreshocks and aftershocks are not a uniform feature of all earthquakes, when occur consecutively, their temporal relationship to the mainshock becomes crucial in determining whether they are classified as a sequence originating from the same epicenter. Furthermore, earthquakes are classified into periodic and geometric. Båth's law explains this by stating that regardless of the mainshock's magnitude, the largest aftershock will have an approximate magnitude difference of 1.2 units [2], [3]. These spatial characteristics align with the temporal characteristics described by the Omega sequences that regards increasing seismicity rate (ISR) as foreshock sequences, where time intervals between events gradually decrease until the mainshock, and decreasing seismicity rate (DSR) as aftershock sequences, where time intervals between events progressively increase, indicating reduced activity. In addition, earthquakes are characterized by their depths and magnitudes.

The depth refers to the distance of the earthquake's point of origin to the Earth's surface. Magnitude, on the other hand, represents the amount of energy released from the rupture's center, with higher magnitudes typically leading to severe damages. The modified mercalli intensity (MMI) scale in Table 1, classifies earthquake magnitudes and their corresponding damage impacts.

Earthquake prediction is nonviable and mostly led to fruitless results [2]–[7], this led researchers to focus on clustering to mine meaningful insights from seismic data. K-means is widely used for clustering, however it is highly sensitive to noise and the selection of initial centers. To mitigate these issues, Shang *et al.* [8] introduced a data field-based variant that incorporates time-event location distance to derive better initial cluster points. While this approach yielded decent results, it did not consistently outperform traditional k-means. Additionally, earthquakes follow a magnitude distribution described by the Gutenberg–Richter law [9], which Shang overlooked by omitting magnitude as a feature. Ultimately, k-means produces regular shaped clusters as well, which does not align with the irregular formations observed in seismic events.

Table 1. MMI scale for earthquake [10], [11]

No	Magnitude	Description	MMI	Common effects
1	1.0–1.9	Micro	I	Microearthquakes. Not felt but recorded by seismographs
2	2.0–2.9	Minor	I	Felt slightly. No building damage
3	3.0–3.9	Slight	II-III	Often felt. Shaking indoor objects, rare damage
4	4.0–4.9	Light	IV-V	Indoor shaking, felt by most. Zero to minimal damage
5	5.0–5.9	Moderate	VI-VII	Commonly felt. Zero to moderate damage buildings
6	6.0–6.9	Strong	VII-IX	Moderate damage, strong shaking in the epicentral area
7	7.0–7.9	Major	>=VIII	Damaged buildings, rails bent
8	8.0–8.9	Great		Major damage to buildings, bridges destroyed
9	9.0–9.9	Extreme		Near-total destruction, severe damage, permanent topography changes

Point-density clustering method has emerged to address high-dimensional and irregular shape clusters. This led some researchers [12], [13] to employ density based spatial clustering of applications with noise (DBSCAN) [14] which relies on two parameters: radius (epsilon) to define the neighborhood size and minimum number of points (MinPts) to form a dense region. While it excels at capturing arbitrary cluster shapes, computational intensity limits its practicality for large datasets. Georgoulas *et al.* [15] introduced seismic mass employ density based spatial clustering of applications with noise (SM-DBSCAN), with improved density calculations, however it retains similar scalability issues with DBSCAN. Campello *et al.* [16] proposed hierarchical density based spatial clustering of applications with noise (HDBSCAN) that automatically determines cluster numbers and improves noise-handling. It uses a distance threshold (epsilon) to compute density and assigns cluster labels based on region stability with a minimum cluster size (MinPts). It effectively identifies clusters with varying densities and shapes. Despite these seemingly promising clustering methods, the challenge has been to bridge the gap between the characteristic models of earthquakes and statistical methods in clustering. Static clustering struggles to address the temporal domain, such as the succession of earthquakes originating from the same epicenter within a specific timeframe, known as seismic sequence that might occur on the same day or spread across different days. This sensitivity to the temporal domain is a distinctive characteristic of earthquakes that distinguishes them from other natural hazards. Spatio temporal employ density based spatial clustering of applications with noise (ST-DBSCAN) [17], a modified version of DBSCAN incorporates the temporal domain and ensures clusters do not merge due to variations in the non-spatial values. However, it faces scalability issues as well.

To address the scalability issues from point-density methods, grid-density clustering has gained attention. It divides the spatial domain into a grid structure and assesses the density within each grid cell. Advanced grid-based iso-density line clustering (AGRID+) [18] is an excellent example of a grid-density clustering for high-dimensional data. It enhances accuracy by considering the lowest element in the clustering results. This method introduces an i-th order neighbor to boost efficiency and is robust in clustering n-dimensional data from spatiotemporal datasets. An improved version, spatiotemporal advanced grid-based clustering (ST-AGRID) [19] introduces partitioning adjustments, distance threshold, and density calculation phases. These modifications translate n-dimensional data into three dimensions: longitude, latitude, and time, facilitating precise partitioning in spatiotemporal clusters. Another version, integrated multi-scale temporal and spatial grid clustering (IMSTAGRID) [20] uses cubed cells as units for representing dimensions and normalizes spatial and temporal features, allowing them to fit into cube-shape grids. It aligns well with seismic data that often requires normalization.

In this paper, we utilized IMSTAGRID which employs cubed cells as units that allows for consistent spatiotemporal division. This technique addresses irregular shapes often observed in earthquake clusters, as well as the temporal aspects, acknowledging that earthquakes from the same epicenter could occur in separate

time frames. Table 2 shows a comparison of previously mentioned clustering methods to IMSTAGRID. This paper demonstrates our approach in seismic sequence clustering, including detailed procedure for exploratory data analysis (EDA), feature engineering, clustering, results evaluation and analysis. Earthquake sequence clustering allows researchers to obtain valuable insights into the evolution of seismic activity to enhance the understanding of the nuanced interplay between the spatial and temporal dynamics in seismic data.

Table 2. Comparison of clustering algorithms based on clustering characteristics

No	Algorithms	Density based	Grid based	Irregular cluster shapes	i-th order neighbor	Density compensation	Spatiotemporal clustering
1	K-Means	-	-	-	-	-	-
2	DBSCAN	✓	-	✓	-	-	-
3	HDBSCAN	✓	-	✓	-	✓	-
2	ST-DBSCAN	✓	-	✓	-	✓	✓
3	AGRID+	-	✓	✓	✓	✓	-

2. METHOD

Figure 1 shows our experiment stages for sequence clustering. These stages include dataset integration, EDA, feature engineering, IMSTAGRID clustering, results evaluation and analysis. This section provides a detailed overview of our data and preprocessing methods to ensure a thorough understanding of our research.

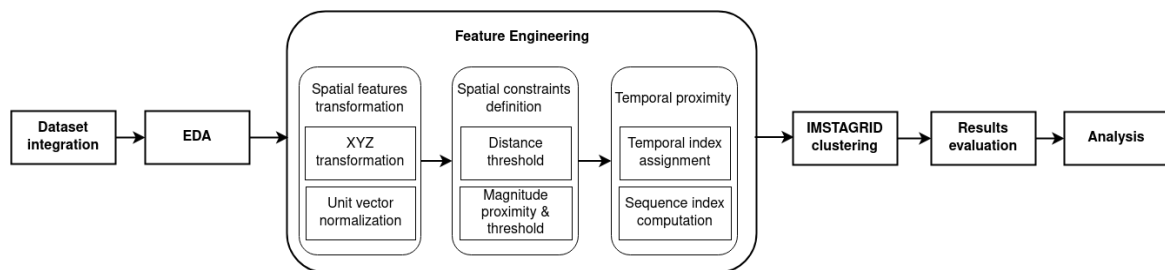


Figure 1. Experiment flow for seismic sequence clustering using IMSTAGRID

2.1. Dataset integration

We sourced two earthquake datasets from the United States Geological Survey (USGS) [21] with periods spanning from 1 January 2017 to 1 January 2023, each comprising 22 attributes as seen in Table 3 [22]. The first dataset represents the conterminous U.S., consisting of 19,843 data which covers a rectangular territory of the U.S., including adjacent areas with latitude and longitude range of (24.6, 50) and (-125, -65). The second dataset is the conterminous Indonesia, comprising 12,826 data that covers a rectangular area with small adjacent territories that share direct borders with Indonesia, the latitude and longitude ranges from (-11.493, 6.433) and (94.036, 141.057). Figures 2(a) and 2(b) shows the geographical regions covered by both datasets.

2.2. Exploratory data analysis

Both datasets are clean. Data distribution in Figures 3(a) and 3(b) demonstrates a significantly right-skewed distribution in the conterminous U.S. dataset. While the conterminous Indonesia dataset is also skewed, it appears milder. A strongly skewed distribution might affect clustering, however, seismic data often contain a wide span of location points, where extreme events are natural occurrences rather than outliers. Given our goal is an optimized sequence clustering considering all spatiotemporal patterns, we ensured real events are not excluded to avoid bias in our findings. Thus, we retain the distributions.

Data plotting shown in Figures 4(a) and 4(b) shows that the Conterminous Indonesia dataset is densely distributed. This is reasonable due to its unique geographical location. Indonesia is located near the meeting point of three major tectonic plates: Eurasian, Indo-Australian, and Pacific Plates that place Indonesia within the deadly "Ring of Fire" [23], which is the world's most seismically active area. Hence, the extremely high earthquake frequency is typical of Indonesia's strong seismic activity.

Table 3. Dataset attributes for the conterminous U.S. and conterminous Indonesia [22]

No	Attribute	Description
1	Time	Time when the earthquake occurred (in ms)
2	Longitude	Degrees east (E) or west (W) of the prime meridian
3	Latitude	Degrees north (N) or south (S) of the equator
4	Depth	Earthquake source depth (in kilometers)
5	Mag	Magnitude of the earthquake
6	magType	Method to calculate earthquake magnitude
7	Nst	Total number of seismic stations used to locate earthquakes
8	Gap	The largest azimuthal distance between neighboring stations
9	Dmin	Epicenter-nearest station horizontal distance (in degrees)
10	Rms	Root-mean-square (earthquake occurrence fit to predicted times)
11	Net	The data contributor's ID
12	Id	An earthquake's unique identification
13	Updated	When the earthquake was most recently updated (in milliseconds)
14	Place	Textual description of the earthquake-affected region (regions/cities)
15	Type	Type of seismic event, in this dataset: earthquake
16	HorizontalError	Uncertainty of reported location of the event (in kilometers)
17	depth Error	Largest earthquake depth error projection (in kilometers)
18	magError	Uncertainty over reported magnitude of the earthquake
19	magNst	Total number of seismic stations utilized to compute the magnitude
20	status	Indicates whether the earthquake has been reviewed by a human
21	LocationSource	The network that first reported the earthquake location
22	magSource	The network that first reported the earthquake magnitude

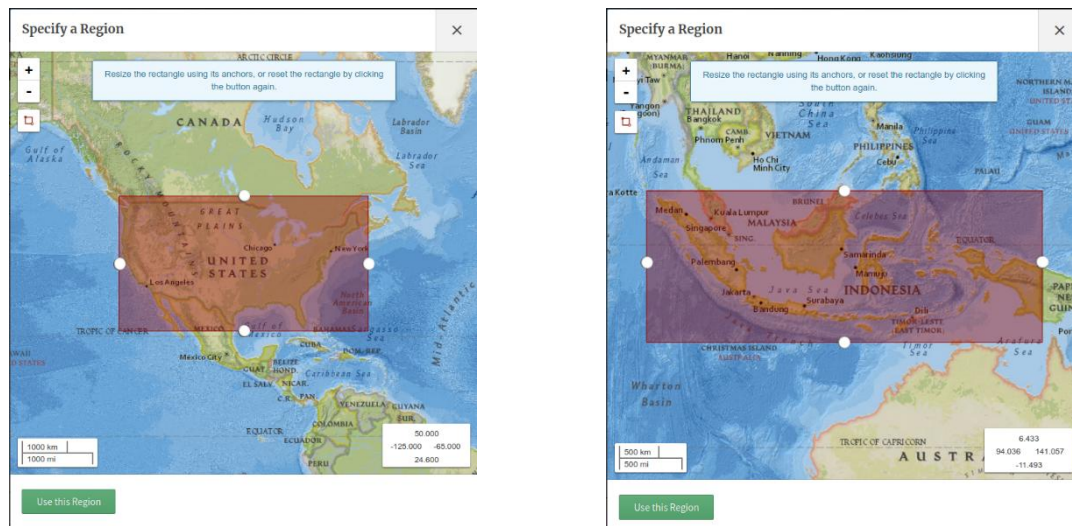


Figure 2. Covered region for (a) the conterminous U.S. dataset and (b) conterminous Indonesia dataset

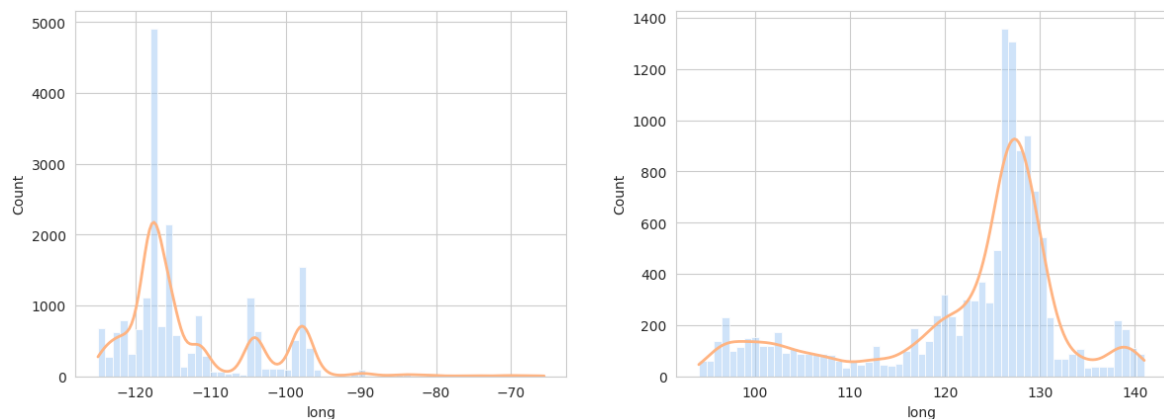


Figure 3. Dataset distribution for (a) the conterminous U.S. and (b) conterminous Indonesia

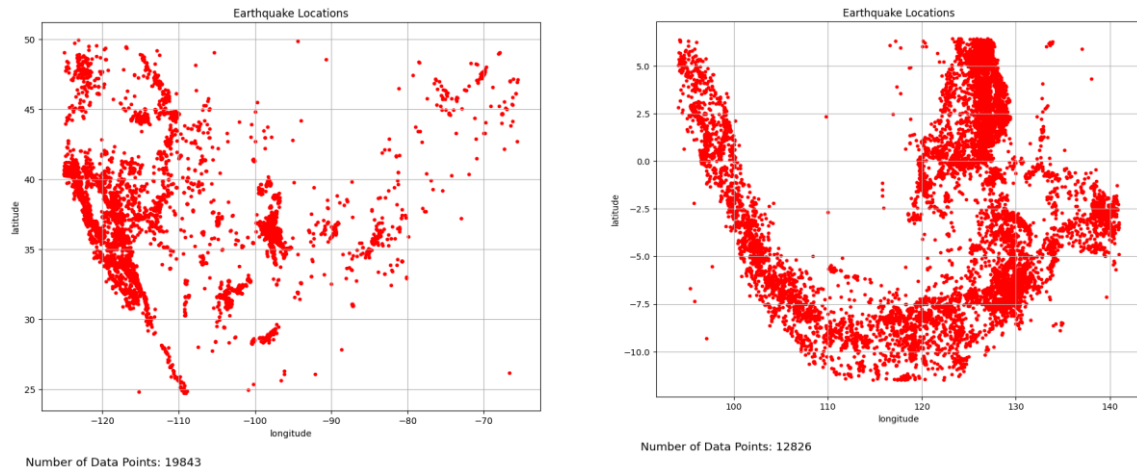


Figure 4. Dataset plotting for (a) the conterminous U.S. and (b) conterminous Indonesia

2.3. Feature engineering

2.3.1. Spatial features translation

Longitude and latitude represent two dimensional attributes in a three-dimensional space. Longitude ranges from -180° to 180° , and latitude from -90° to 90° . This range difference introduces unique challenges in normalization, such as, the circular nature of longitude coordinates implies that the two most extreme values are adjacent instead. Normalization using min-max is not ideal as it leads to distorted spatial relationships where the distance between normalized points does not correspond to real-world distances. Additionally, the physical distance given by longitude varies with latitude, a factor that min-max overlooks, worsening the distortion. To address this, we introduced XYZ transformation with unit vector normalization.

a) XYZ transformation

We transform longitude and latitude to a three-dimensional Cartesian coordinate system with the origin at the center mass of the earth, following [23]:

$$\lambda = \text{longitude (radians)}$$

$$\theta = \text{latitude (radians)}$$

$$R = \text{Earth radius (6,371 kilometers)}$$

$$X = R \cdot \cos(\theta) * \cos(\lambda) \quad (1)$$

$$Y = R \cdot \cos(\theta) * \sin(\lambda) \quad (2)$$

$$Z = R \cdot \sin(\theta) \quad (3)$$

As depicted in Figure 5, the X and Y axes from (1) and (2) span the equatorial plane, while the Z axis from (3) corresponds to the rotating axis.

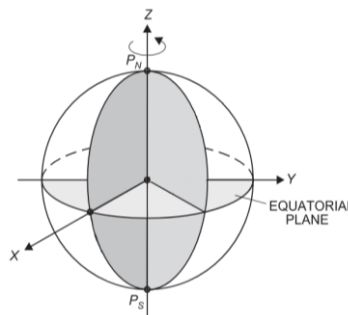


Figure 5. Astronomical equatorial system, simplified from Torge *et al.* [24]

b) Unit vector normalization

Using the obtained XYZ values, we performed unit vector normalization. Unlike min-max scaling which scales features to a range between 0 to 1, unit vector normalization as seen in (4) to (7) scales coordinates by their magnitude, preserving angles and geographical directions on a three-dimensional space.

$$XYZ \text{ magnitude} = \sqrt{X^2 + Y^2 + Z^2} \quad (4)$$

$$X \text{ normalized} = \frac{X}{\text{magnitude}} \quad (5)$$

$$Y \text{ normalized} = \frac{Y}{\text{magnitude}} \quad (6)$$

$$Z \text{ normalized} = \frac{Z}{\text{magnitude}} \quad (7)$$

Figures 6(a) and 6(b) demonstrate the visualization of XYZ coordinates for the conterminous U.S and the conterminous Indonesia dataset. We proceed with the equatorial span (X, Y) as spatial features for clustering.

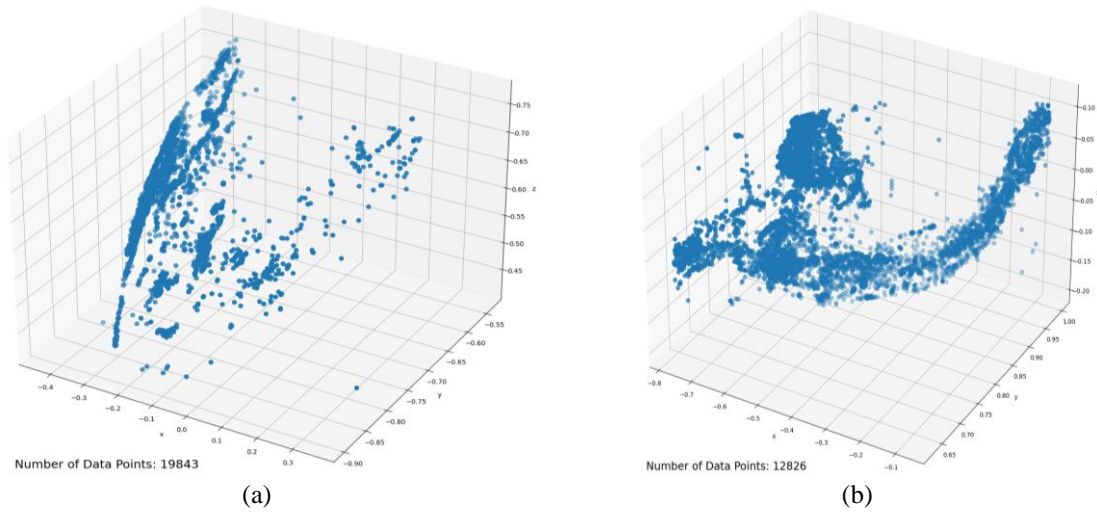


Figure 6. XYZ coordinates in (a) conterminous U.S. and (b) conterminous Indonesia

2.3.2. Spatial constraints

Determining whether an event is an aftershock based on its spatial distance to adjacent events is oversimplification. Aftershocks are heavily influenced by other factors, such as magnitude. We integrated a few constraints for sequence clustering: distance threshold, magnitude proximity and magnitude threshold.

a) Distance threshold

Longitude and latitude thresholds, referred to as distance threshold d from here onward, are crucial to determine if an event is part of a seismic sequence from a similar mainshock. Aftershocks typically occur in close proximity to the mainshock, a common guideline is that aftershocks tend to occur approximately 100 km from its mainshock [22]. Computing d using planar approximation is not ideal because it assumes a flat surface, ignoring the Earth's curvature. To address this, we implemented the Haversine formula as seen from (8) to (10), which assumes an approximately spherical Earth to account for its curvature:

$$\lambda = \text{longitude (radians)}$$

$$\theta = \text{latitude (radians)}$$

$$R = \text{Earth radius (approx. 6,371 kilometers)}$$

$$d = \text{distance threshold (kilometers)}$$

$$a = \sin^2 \left(\frac{\Delta\lambda}{2} \right) \cos \lambda_1 \cdot \cos \lambda_2 * \sin 2 \left(\frac{\Delta\theta}{2} \right) \quad (8)$$

$$c = 2 * \operatorname{atan} 2 (\sqrt{a}, \sqrt{1-a}) \quad (9)$$

$$d = R * c \quad (10)$$

Using the distance threshold, we temporarily classify earthquakes on the same day within the radius of 100 km as a single seismic sequence. Otherwise, they are treated as distinct events. Each cluster will be further filtered through magnitude proximity and temporal threshold.

b) Magnitude proximity

Magnitude directly affects the aftershock distribution. While there are no widely accepted standards defining earthquake radius coverage based on its magnitude, we compiled a magnitude proximity M_p of seismic events based on the MMI Scale in Table 1.

1.0-4.9 (minor) : 100 km radius influence

5.0-5.9 (moderate): 500 km radius influence

6.0-9.9 (strong) : 1,000 km radius influence

Knowing that the largest aftershock will have an approximate magnitude threshold M_t of 1.2 units [2]–[3] from its mainshock, we assess distance threshold, magnitude proximity and magnitude threshold to determine if an event is an aftershock. For events occurring within the same timeframe (day), let D_{AB} represent the distance between Event A, denoted as $A(M_A)$ with magnitude M_A , and Event B, denoted as $B(M_B)$ with magnitude M_B . Additionally, define R as 100 km radius threshold and $M_{thresh}=M_B \leq (M_A-1.2)$ as the magnitude threshold M_t for an event to be categorized as an aftershock. The categorization Event B is:

If $D_{AB} < R$, then Event B is an aftershock regardless of M_B

If $D_{AB} > R$ and $M_{thresh}=false$, then Event B is a distinct event

If $M_{thresh}=true$, check these conditions:

If $M_A \leq 4.9$, then $R_A=100$ km. If $D_{AB} < R_A$, then Event B is an aftershock

If $5.0 \leq M_A \leq 5.9$, then $R_A=500$ km. If $D_{AB} < R_A$, then Event B is an aftershock

If $M_A \geq 6.0$, then $R_A=1,000$ km. If $D_{AB} < R_A$, then Event B is an aftershock

In this framework, Event A is considered as the mainshock and Event B is determined by both the distance threshold criterion ($D_{AB} < R_A$) and magnitude threshold criterion ($M_{thresh}=true$). Suppose, A (7.0) and B (6.0) are $D_{AB}=120$ km apart. $D_{AB} > R_A$, indicating Event B as a distinct event since it does not meet the distance threshold criterion. However, considering the influence radius of Event A ($R_A=1000$ km) for $M_A=7.0$ and $\Delta M_{AB}=1.0$, making $M_{thresh}=true$, we adjust the categorization to classify Event B as an aftershock of Event A, aligning it with the magnitude threshold M_t criterion.

2.3.3. Temporal threshold

Earthquakes exhibit a unique sensitivity to temporal dimension that results in events originating from the same epicenter occurring successively within different timeframes. To capture these unique characteristics, we introduce new key features: *temporal_idx* and *sequence_idx*. Each of these features is discussed below.

a) Temporal index assignment

Let E be the set of earthquake events, each associated with timestamp t_i . We define a function $f_{temporal_idx}(t_i)$ that assign a unique identifier $f_{temporal_idx}$ as seen in (11) to each event based on timestamp t_i :

$$t_{temporal_idx} = f_{temporal_idx}(t_i) \quad (11)$$

This allows events occurring on the same timeframe to be assigned with the same $t_{temporal_idx}$. We tested two temporal timeframes, which are daily-based and weekly-based. For daily-based sequence clustering, $t_{temporal_idx}$ is for events occurring on the same day. For weekly-based clustering, we group events into a set of seven neighboring $t_{temporal_idx}$. As seen in (12), let S_{week} represent the set of weekly clusters:

$$S_{week} = \{S_1, S_2, \dots, S_n\} \quad (12)$$

Each S_i contains events with consecutive $t_{temporal_idx}$ values over a seven-day period.

b) Sequence index computation

Among events sharing the same $t_{temporal_idx}$, we assess the distance threshold d , magnitude proximity M_p , and magnitude threshold M_t . Then we group sequence index as follows:

$$t_{sequence_idx} = \{t_{temporal_idx} + 1 \text{ similar events exist} \mid t_{temporal_idx} \text{ no similar events exist}\} \quad (13)$$

From (13), $t_{sequence_idx}$ is assigned based on the absence of presence of similar events with the same $t_{temporal_idx}$ and satisfies M_p and M_t . Events classified as part of a single sequence share similar $t_{sequence_idx}$ and $t_{temporal_idx}$, indicating there are events that share similar $t_{temporal_idx}$, but not part of the same sequence.

2.4. IMSTAGRID clustering

When determining distance threshold r to obtain neighboring points and neighboring cells, uneven partitioning leads to gaps as seen in Figure 7. IMSTAGRID [20] maintains a uniform interval value L that generates matching values for spatial and temporal dimensions, allowing for a cube shape.

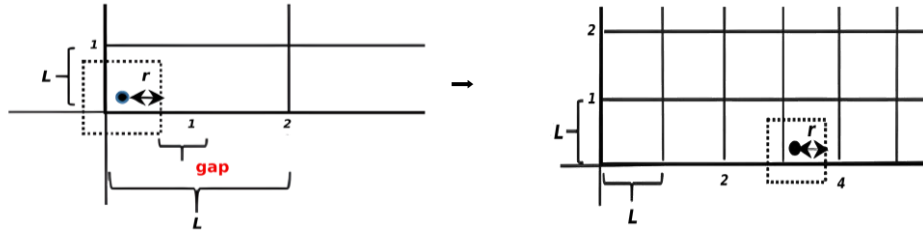


Figure 7. IMSTAGRID addressed the gap from uneven partition using a uniform L value in spatial and temporal dimensions to determine the distance threshold r . [20]

Distance threshold r to determine density threshold is obtained and used to compute density compensation $C_{densities}$ as shown in (14):

$$C_{densities} = densities(O_i) * \frac{volume\ of\ O_i\ cube}{volume\ of\ all\ O_i\ neighboring\ cubes} \quad (14)$$

The *volume of all O_i neighboring cubes* is computed based on order of proximity with the i -th neighbor.

Clustering process shown in Figure 8 includes a point α on the top right corner of $C\alpha$ cube, implies only that cube will be included in clustering. The *volume of all O_i neighboring cubes* is calculated considering four neighbors where the volume must be determined. To account for the unique temporal characteristics tied to seismic data, we incorporated the methods proposed in point 2.3.3 into IMSTAGRID.

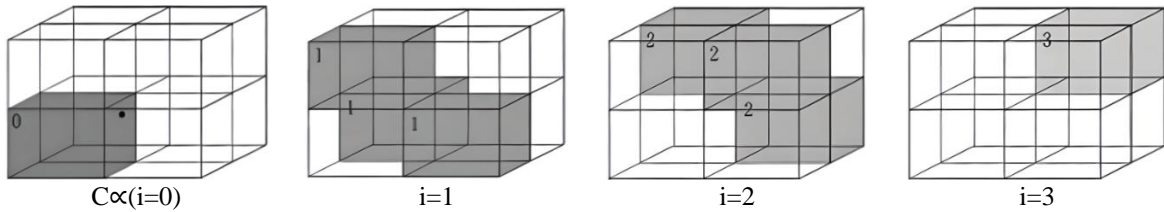


Figure 8. The i -th neighbor order of an α point of a $C\alpha$ cube [20]

2.5. Evaluation metric

To evaluate and gain insights into the spatiotemporal patterns of earthquake sequences, we chose a quantitative metric which is the Silhouette score. It provides a way to quantify the degree of intra-cluster cohesion and inter-cluster separation on a scale of -1 to 1, with metric maximization close to 1 is preferred. The Silhouette value for each seismic event i in a cluster C_i is determined as seen in (15) [25]:

$$ss(i) = \frac{B-A}{\max(AB)} \quad (15)$$

Where $A(i)$, as seen in (16) represents the mean dissimilarity between the seismic event i and other events in the same cluster. The intra-cluster cohesion indicates how closely events within a cluster are related in terms of their spatiotemporal features. $d(i, j)$ is the distance metric between events i and j .

$$A(i) = \frac{1}{|C_l|-1} \sum_{k \in C_l, i \neq j} d(i, j) \quad (16)$$

Each $B(i)$ as seen in (17) denotes the minimum average dissimilarity between seismic event i and the events in a different cluster, with the selection of the cluster aimed at minimizing this dissimilarity. The inter-cluster separation, signifying how distinct one cluster is from its neighboring clusters.

$$B(i) = \min \frac{1}{|C_o|} \sum_{k \in C_o} d(i, j) \quad (17)$$

A Silhouette score close to 1 indicates well-separated seismic sequence clusters where events are highly similar and distinct from other sequences. A score near 0 suggests potential overlap in separation and a score close to -1 implies inadequate separation, possibly misallocated earthquake events.

3. RESULTS AND DISCUSSION

We attempted to cluster earthquake events into two temporal timeframes: daily-based and weekly-based. We thoroughly investigated various $\lambda = [0.01, 0.05, 0.1, 0.2, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$ and $\theta = [0.1, 0.2, 0.5, 1, 2, 5, 10]$ values by looping through them in order to find the optimal combination for each temporal aggregate. First, we evaluated the conterminous U.S. clustering results as shown in Figure 9(a) and Figure 9(b). A Silhouette score of 0.935 from 83 clusters shows exceptionally well separated clusters in daily-based clustering using $\lambda=0.1$ and $\theta=1$. It implies that earthquakes within each cluster are highly similar while being notably distinct from events in other clusters. The formation of 83 clusters indicates that earthquake sequence varies greatly throughout the day and corresponds to certain time intervals. In weekly-based clustering, we achieved a Silhouette score of 0.755 with $\lambda=0.01$ and $\theta=1$. This suggests well-separated sequence clusters where earthquakes in each of the 30 clusters are comparable yet unique from other clusters. The formation of 30 clusters reveals that seismic activity varies throughout the week, representing diverse temporal patterns on a larger time scale.

The conterminous Indonesia dataset is densely distributed due to the inherently high seismic activity in the area. While density may not directly correlate with silhouette values, it highlights the challenges in capturing complex temporal patterns in seismic activity. Figure 10(a) shows that daily-based clustering using a combination of $\lambda=0.5$ and $\theta=2$, yielded a silhouette value of 0.782 from 983 clusters implying strongly well-separated sequence clusters. Given the high earthquake frequency in Indonesia, the development of 983 clusters from 12,826 data points depicts a fine temporal granularity with a level of detail that can capture complicated temporal patterns. In the weekly-based clustering as shown in Figure 10(b), a combination of $\lambda=0.01$ and $\theta=1$, yielded a Silhouette score of 0.610 from 118 clusters. Although lower than the daily-cluster, it still shows relatively well-separated clusters that are comparable yet distinct from occurrences in other clusters.

Table 4 presents the results of IMSTAGRID clustering for both datasets. In the conterminous U.S. dataset, cluster count decreases from 83 in daily clustering to 30 in weekly clustering. To gain a deeper understanding of this reduction, Figure 3 offers a visualization of the dataset, revealing a skewed distribution where certain data points are heavily concentrated on one side while others appear sparser. This distribution signifies an uneven distribution of seismic activity across time, a characteristic frequently encountered in real-world seismic datasets. In contrast, the reduction from 983 daily clusters to 118 weekly clusters in the conterminous Indonesia dataset appears to be a more logical shift aimed at incorporating broader temporal patterns. As depicted in Figure 3, this dataset displays an exceptionally dense distribution, indicating that seismic events are spread out densely across time. While the Silhouette values obtained are reasonable, the relatively lower scores could be attributed to factors such as the inherent variability in seismic activity, which can lead to overlapping clusters in highly dense distribution and reduce Silhouette values.

Our technique presented a remarkable efficacy in clustering spatiotemporal seismic sequences and revealed noteworthy trends across both datasets where the optimal values of λ for weekly basis clustering tends to be lower than those for daily basis clustering across all datasets. This observation leads us to derive meaningful insights in relation to the temporal domain, outlined below:

- Sensitivity to temporal granularity. Lambda λ choice can be sensitive to temporal granularity. Daily clustering, being more fine-grained, requires a higher λ to account for higher event frequency. On the other hand, weekly clustering that aggregates data over a broader time might benefit from a lower λ to capture larger temporal patterns.
- Complex temporal dynamics. Seismic data are known to carry complex temporal dynamics, including short and long occurrence patterns. Daily clustering captures rapid patterns, while weekly clustering focuses on extended patterns. Lambda λ choice reflects the need to balance temporal patterns.

Unlike traditional clustering methods that focus on clustering spatial data, our approach dynamically captures earthquake occurrences overtime, providing information about the evolving nature of seismic events in relation to the temporal domain. Our method enabled the identification of patterns that are overlooked in static clustering methods, where fine-grained daily clustering requires higher λ to capture rapid fluctuations and weekly clustering benefits from lower λ to cover broader trends. However, further study might enhance our clustering algorithms to capture new characteristics and increase accuracy. Such as, identifying seismic sequences with mixed temporal features. Ultimately, automated techniques for identifying the optimal λ and θ values remain a future research focus as well to enhance repeatability in seismic sequence analysis.

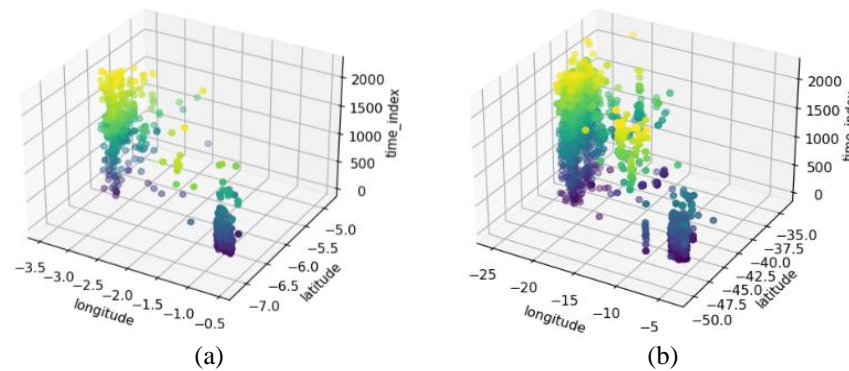


Figure 9. The conterminous U.S. dataset with Silhouette scores of 0.935 and 0.755, respectively (a) daily and (b) weekly clustering

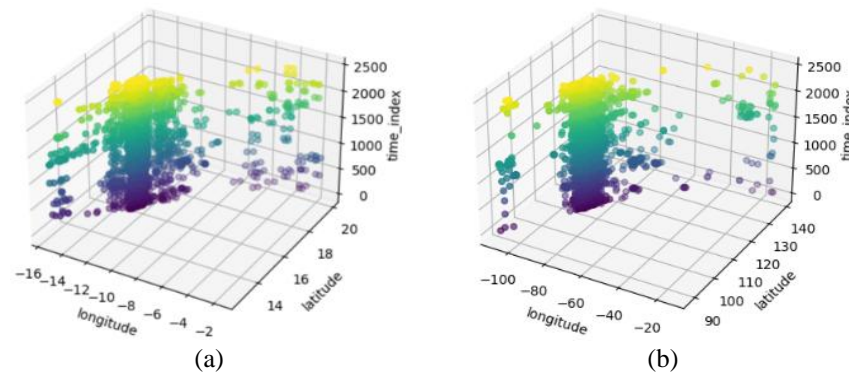


Figure 10. The conterminous Indonesia dataset with silhouette scores of 0.782 and 0.610, respectively (a) daily and (b) weekly clustering

Table 4. IMSTAGRID seismic sequence clustering results across two datasets

No	Dataset	Points data	Temporal aggregate	Best λ	Best θ	Silhouette score	Clusters count
1	Conterminous U.S.	19,843	1	0.1	1	0.935	83
			7	0.01	1	0.755	30
2	Conterminous Indonesia	12,826	1	0.5	2	0.782	983
			7	0.01	1	0.61	118

4. CONCLUSION

Our cube-based clustering technique effectively clusters earthquake sequences over time, this approach differs from traditional clustering methods that focus on static data points. We explored IMSTAGRID by incorporating unique temporal characteristics of earthquake sequences. Our experiments revealed significant patterns in selecting λ for temporal analysis. We discovered that the optimum λ values for weekly clustering consistently shifted lower than values for daily clustering across two datasets. Daily clustering of the conterminous U.S. dataset obtained 83 clusters with a Silhouette score of 0.935, while weekly clustering produced 30 clusters with a Silhouette score of 0.755. In the conterminous Indonesia dataset, daily clustering

produced 983 clusters with a Silhouette score of 0.782, while weekly clustering produced 118 clusters with a Silhouette score of 0.610. These findings provide a niche understanding of the temporal evolution of seismic activity that emphasizes the importance of adapting λ to temporal granularity. Fine-grained daily clustering benefits from higher λ values to catch rapid fluctuations, while weekly clustering excels with lower λ values to capture wider temporal patterns. Nevertheless, it is critical to acknowledge the unpredictability in seismic data, which might result in overlapping clusters, for instance as we observed in the conterminous Indonesia dataset that led to lower Silhouette values. This highlights the complexities of seismic data with patterns beyond clustering algorithms reach.

ACKNOWLEDGEMENT

This work is supported by the Research and Technology Transfer Office, Bina Nusantara University as a part of Bina Nusantara University's International Research Grant entitled Automatic Determination of Spatial and Temporal Dimension Interval Values in Grid and Density-Based Clustering Algorithms with contract number: 029/VRRTT/III/2023 and contract date: March 1, 2023.




REFERENCES

- [1] V. Svalova, *Earthquakes - Forecast, Prognosis and Earthquake Resistant Construction*, London, United Kingdom: IntechOpen, 2018, doi: 10.5772/intechopen.71298.
- [2] J. Žalohar, *The omega-theory: a new physics of earthquakes*. Amsterdam, Netherlands: Elsevier Science, 2018.
- [3] Udías and E. Buforn, *Principles of Seismology*. Cambridge, United Kingdom: Cambridge University Press, 2017, doi: 10.1017/9781316481615.
- [4] H. Wang *et al.*, "Pre-earthquake observations and their application in earthquake prediction in China," *Pre-Earthquake Processes: A Multidisciplinary Approach to Earthquake Prediction Studies*, pp. 19–39, 2018, doi: 10.1002/9781119156949.ch3.
- [5] G. Martinelli, "Contributions to a history of earthquake prediction research," *Geophysical Monograph Series*, vol. 234, pp. 67–76, 2018, doi: 10.1002/9781119156949.ch5.
- [6] J. B. Rundle, D. L. Turcotte, R. Shcherbakov, W. Klein, and C. Sammis, "Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems," *Reviews of Geophysics*, vol. 41, no. 4, 2003, doi: 10.1029/2003RG000135.
- [7] F. Huang *et al.*, "Studies on earthquake precursors in China: a review for recent 50 years," *Geodesy and Geodynamics*, vol. 8, no. 1, pp. 1–12, 2017, doi: 10.1016/j.geog.2016.12.002.
- [8] X. Shang, X. Li, A. Morales-Esteban, G. Asencio-Cortés, and Z. Wang, "Data field-based K-means clustering for spatio-temporal seismicity analysis and hazard assessment," *Remote Sensing*, vol. 10, no. 3, 2018, doi: 10.3390/rs10030461.
- [9] Y. Zhang, D. Zhou, J. Fan, W. Marzocchi, Y. Ashkenazy, and S. Havlin, "Improved earthquake aftershocks forecasting model based on long-term memory," *New Journal of Physics*, vol. 23, no. 4, Apr. 2021, doi: 10.1088/1367-2630/abeb46.
- [10] G. J. V. Wijngaarden, and H. Kars, "Long-term effect of seismic activities on archaeological remains: a test study from Zakynthos, Greece," in *Ancient Earthquakes*, Geological Society of America, 2010, pp. 145–156, doi: 10.1130/2010.2471(13).
- [11] Earthquake Hazards Program, "The modified mercalli intensity scale," *USGS Science for a Changing World*, 2023. Accessed: Mar. 8, 2023. [Online]. Available: <https://www.usgs.gov/programs/earthquake-hazards/modified-mercalli-intensity-scale>
- [12] S. Scitovski, "A density-based clustering algorithm for earthquake zoning," *Computers & Geosciences*, vol. 110, Jan. 2018, doi: 10.1016/j.cageo.2017.08.014.
- [13] M. Kazemi-Beydokhti, R. Ali Abbaspour, and M. Mojarab, "Spatio-temporal modeling of seismic provinces of Iran using DBSCAN algorithm," *Pure and Applied Geophysics*, vol. 174, no. 5, May 2017, doi: 10.1007/s00024-017-1507-0.
- [14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," *KDD-96 Proceedings*, pp. 226–231, 1996.
- [15] G. Georgoulas, A. Konstantaras, E. Katsifarakis, C. D. Stylios, E. Maravelakis, and G. J. Vachtsevanos, "Seismic-mass' density-based algorithm for spatio-temporal clustering," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4183–4189, 2013, doi: 10.1016/j.eswa.2013.01.028.
- [16] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," *Advances in Knowledge Discovery and Data Mining*, 2013, pp. 160–172, doi: 10.1007/978-3-642-37456-2_14.
- [17] D. Birant and A. Kut, "ST-DBSCAN: an algorithm for clustering spatial-temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, Jan. 2007, doi: 10.1016/j.datak.2006.01.013.
- [18] D. Fitrihanah, A. N. Hidayanto, H. Fahmi, J. L. Gaol, and A. M. Arymurthy, "ST-AGRID: A spatio temporal grid density-based clustering and its application for determining the potential fishing zones," *International Journal of Software Engineering and its Applications*, vol. 9, no. 1, pp. 13–26, 2015, doi: 10.14257/ijseia.2015.9.1.02.
- [19] Y. Zhao, J. Cao, C. Zhang, and S. Zhang, "Enhancing grid-density based clustering for high dimensional data," *Journal of Systems and Software*, vol. 84, no. 9, pp. 1524–1539, Sep. 2011, doi: 10.1016/j.jss.2011.02.047.
- [20] D. Fitrihanah, H. Fahmi, A. N. Hidayanto, and A. M. Arymurthy, "Improved partitioning technique for density cube-based spatio-temporal clustering method," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, Nov. 2022, doi: 10.1016/j.jksuci.2022.08.006.
- [21] USGS, "Search earthquake catalog," *USGS Science for a Changing World*, 2023. Accessed: Jan. 5, 2023. [Online]. Available: <https://earthquake.usgs.gov/earthquakes/search/>
- [22] USGS, "ANSS comprehensive earthquake catalog (ComCat) documentation," *USGS Science for a Changing World*, 2023. Accessed: Jan. 5, 2023. [Online]. Available: <https://earthquake.usgs.gov/data/comcat/>
- [23] J. Bell, *The earth book: from the beginning to the end of our planet, 250 milestones in the history of earth science*. New York: Union Square+ ORM, 2019.
- [24] W. Torge, J. Müller, and R. Pail, *Geodesy*. Berlin, Germany: De Gruyter Textbook, 2023, doi: 10.1515/9783110723304.
- [25] M. Mollaiian, G. Dörgö, and A. Palazoglu, "Performing multi-objective optimization alongside dimension reduction to determine number of clusters," *Processes*, vol. 10, no. 5, May 2022, doi: 10.3390/pr10050893.




BIOGRAPHIES OF AUTHORS

Silviya Hasana    holds a Master of Engineering degree from the Nara Institute of Technology, Japan, which she completed in 2019. Previously, she earned her undergraduate degree in Computer Engineering from Sebelas Maret University, Indonesia. She has worked as a Computer Science lecturer at Kalbis Institute, Indonesia and subsequently, at Binus University, Indonesia. Currently, she is working as a data engineer in Indonesia. Her main research fields are in artificial intelligence, including machine learning, deep learning, and data mining with interests in vision research, healthcare and disaster management. She can be contacted at email: silviahasana@gmail.com.






Wina Permana Sari    holds a Master's Degree in Computer Science from the University of Indonesia, Depok, Indonesia in 2016. She also received her S.T (Informatics Engineering) degree from Telkom University, Bandung, Indonesia in 2013. She worked as a lecturer at the Computer Science department at Kalbe Institute in 2016. She is currently a permanent lecturer at the School of Computer Science at Bina Nusantara University, Malang, Indonesia from 2018. She has also been active at SINAM LAB Bina Nusantara University. Her research includes data mining, text mining, e-government, information systems, and knowledge management. She also has published several papers in international conferences and journals since July 2013. She can be contacted at email: wina.sari001@binus.ac.id.



Rojali    holds a Doctor of Computer Science degree from Bina Nusantara University, Indonesia, with the Dissertation "Increasing number of hidden bit in steganography using 25 pixel-value-differencing (PVD)". He also received his B.Sc. (Mathematics) from Bina Nusantara University. He also received his M.Sc. in Computer Science from IPB University. He is currently lecturing with the Master of Computer Science at Bina Nusantara University, Indonesia. His research interests are in cryptography, steganography, numerical analysis or scientific computing. He can be contacted at email: rojali@binus.edu.



Devi Fitrianah    is a lecturer and researcher at the Master of Computer Science Department at Bina Nusantara University. She received her Bachelor's degree in Computer Science from Bina Nusantara University in 2000 and a Master's degree in Information Technology and a Ph.D. degree in Computer Science from Universitas Indonesia in 2008 and 2015 respectively. She was rewarded with a sandwich program at the Laboratory for Pattern Recognition and Image Processing and GIS (PRIPGIS Lab) Department of Computer Science, Michigan State University, East Lansing, Michigan, USA in 2014. She can be contacted at email: devi.fitrianah@binus.edu.