CLUSTERING AND RISK ANALYSIS OF THE EARTHQUAKE IN SULAWESI USING MINI BATCH K-MEANS, K-MEDOIDS, AND MAXIMUM LIKELIHOOD METHOD

Amirin Kusmiran

Department of Physics, Faculty of Science and Technology, Universitas Islam Negeri Alauddin Makassar, Makassar, Indonesia, amirin.kusmiran@uin-alauddin.ac.id

Email Correspondence : amirin.kusmiran@uin-alauddin.ac.id

Abstract: The earthquake events have been widely analyzed using a statistical approach. Therefore, the sole purpose of this research is clustering and risk analysis of earthquake events based on the combination of machine learning and statistics. The machine learning, conducted by Mini Batch K-Means and K-Medoids, is validated by the Davies-Bouldin index method to earthquake events cluster. Furthermore, the statistics approach conducted by the maximum likelihood method is to estimate the *b*-value and *a*-value of earthquake events. The data used in the earthquake events analysis in Sulawesi have a magnitude ≥ 5 SR during the period 1980-2022. The results show that the Mini Batch K-Means method is more efficient and accurate than the K-Medoids, and can cluster the earthquakes, namely cluster 0 below 100 km (shallow earthquake), cluster 1 above 100 km to 350 km (medium earthquake), cluster 2 above 350 km (deep earthquake), while K-Medoids method has two clusters namely cluster 0 below 100 km (shallow earthquake), and cluster 1 above 100 km to 350 km. The regions with *b*-value and *a*-value less than 0.9 and 7.5, respectively, and in cluster 0, namely the western part of North Sulawesi, Gorontalo, Middle Sulawesi, and West Sulawesi Province, are as vulnerable to earthquake disasters. Meanwhile, the region in cluster 1 and cluster 2 with *b*-value and *a*-value more than 0.9 and 7.5 respectively namely South Sulawesi, the Northern part of North Sulawesi, and Southeast Sulawesi Province, are categorized as minor earthquake disasters. Furthermore, the clustering and risk analysis based on these methods results are good performance, which has recognised cluster and vulnerability of the earthquake events.
Keywords: Mini Batch K-Means; K-Medoids; *a*-value; *b*-value; Seismicity

Abstrak: Kejadian gempa bumi telah banyak dianalisis dengan menggunakan pendekatan statistik. Oleh karena itu, tujuan penelitian untuk menganalisis kejadian gempa dengan menggunakan kombinasi pendekatan *machine learning* dengan statistik. Pendekatan *machine learning* dilakukan dengan metode baru yakni metode Mini Batch K-Means dan K-Medoids yang divalidasi dengan metode Davies-Bouldin indeks yang digunakan untuk mengklaster kejadian gempa, sedangkan pendekatan secara statistik dilakukan dengan metode *maximum likelihood* untuk mengestimasi kerentanan gempa bumi berdasarkan nilai-*b* dan nilai-*a*. Data yang digunakan yakni data kejadian gempa di Sulawesi dengan magnitudo ≥ 5 SR dengan periode 1980-2022. Hasil menunjukan bahwa metode Mini Batch K-Means lebih effisien dan akurat dibandingkan dengan metode K-Medoids, dan mengklasifikasi tiga klaster kedalaman gempa, yakni klaster 0 dengan kedalaman kurang dari 100 km (gempa dangkal), klaster 1 dengan kedalaman

diantara 100 km dengan 350 km (gempa menengah), klaster 2 dengan kedalaman lebih dari 350 km (gempa dalam). Sementara metode K-Medoids dua klaster kedalama gempa, yakni klaster 0 dengan kedalaman dibawah 100 km (gempa dangkal), dan klaster 1 dengan kedalaman lebih dari 100 km. Beberapa wilayah yang mempunyai nilai-*b* dan nilai-*a* secara berurutan kurang dari 0,9 dan 7,5 dan termasuk ke dalam klaster 0, yakni Provinsi Sulawesi Utara bagian barat, Gorontalo, Sulawesi Tengah, dan Sulawesi Barat dikategorikan rawan terhadap bencana gempa; Sedangkan wilayah yang termasuk ke dalam klaster 1 dan klaster 2 dengan nilai-*b* dan nilai-*a* secara berurutan lebih dari 0,9 dan 7,5 yakni Provinsi Sulawesi Selatan, Sulawesi Utara bagian Utara, dan Sulawesi Tenggara dikategorikan sebagai rendah terhadap bencana gempa. Dengan demikian, kedua metode dapat digunakan untuk meng-klaster gempa dan identifikasi kerentanan kejadian gempa bumi.

Kata kunci: Mini Batch K-Means; K-Medoids; nilai-*a*; nilai-*b*; Seismisitas.

Introduction

Sulawesi is a triple junction of Eurasian, Hindia-Australia, and Pacific Ocean plates so Sulawesi has complex structural geology such as the Palu-Koro fault, North Sulawesi trench, Mantano fault, East Sangihe trust, Tolo trust, Batui trust, Lawanopo Fault, Kolaka fault, Waianae fault, and Tukang Besi (White et al., 2017). The effects of that collisions are the release of seismic waves in the subduction zone where the energy is spread out radially in the subsurface of has caused earthquakes significantly. The energy effect of the seismic waves can cause changes in stress in the near epicenter of the earthquakes. The accumulation of this stress change can cause changes in seismic activity around the post-earthquake area so that it can trigger earthquakes in the future. The relationship between seismic stress or recurrence of the earthquake with time failure is divided into interseismic, coseismic, and postseismic (Nurana et al., 2021). The stress accumulation of earthquakes event and ductility of rock correlated with the *b*-value and *a*-value parameters. The b-value indicates the stress of the tectonic parameter, and the *a*-value indicates seismicity within a given region (Scholz, 1968).

The a value and b value parameter have been widely used for recurrence earthquake identification, such as spatial-temporal condition recent seismicity in Sumatera (Nurana et al., 2021), *a*-value and *b*-value indicator of the earthquake potential in West Nusa Tenggara (Ernandi & Madlazim, 2020), statistical analysis of South Sulawesi earthquake (Kurniawati & Muhammad Rahman, 2020), and *b*-value and seismic hazard investigation (Ray et al., 2019). Based on their research of the *a*-value and *b*-value, we found a difference in characteristic seismicity within the region. Consequently, stress states are also varied within the area because of the material characteristic of the subsurface difference (Scholz, 1968,

Elkawnie: Journal of Islamic Science and Technology Vol. 9, No. 1, June 2023
(www.jurnal.ar-raniry.ac.id/index.php/elkawnie)

DOI: 10.22373/ekw.v9i1.13027 **| 2**

2015). The relations of the frequency earthquake event with the *a*-value and the *b*-value parameter were revealed by Gutenberg & Richter (1944). Furthermore, the comprehensive interpretation, of machine learning is recently used to analyze the event earthquake characteristics where the resemblance characteristics are signed by equal clusters.

Clustering is one of the parts of data mining that is used for exploratory information on seismic events and is part of unsupervised learning categories. From these techniques, the earthquake or seismic patterns of the unlabeled attribute can be determined. The gaussian mixture model has been applied by Seydoux et al. (2020) to clustering earthquake signals and this method is one unsupervised machine learning technique. Inherently, K-Medoids and K-Means clustering is also applied to clustering earthquake risk where R programming is implemented in the clustering process (Ahmad & Irsalinda, 2020; Novinanti et al., 2017; Rifa et al., 2020; Yuan, 2021). The euclidean distance used to find the centroid cluster of the catalog earthquake is the square of the straight line distance of two points in Euclidean space.

Some research has been conducted in the study area to explain the tectonic process in Sulawesi, such as tectonic activities in Sulawesi (Zakaria & Sidarto, 2015), geological history of Western Sulawesi (White et al., 2017), seismic hazard assessment (Cipta et al., 2016; Eva, 2016). Clustering earthquake risk by K-Means and K-Medoids method with R programming has been conducted by Ahmad & Irsalinda (2020), Novinanti et al. (2017), and Rifa et al. (2020). Moreover, research on seismic activities must be continuously conducted to explain the recent seismic activities in both spatial and temporal conditions. For seismic activities analysis, the raw data is split to be some attributes such as time, date, magnitude, and depth. Then, the maximum likelihood method is applied to determine the *b*-value of the Gutenberg-Richter formula, and Mini Batch K-Means of machine learning methods are applied to describe the earthquake event clusters, where this method is not applied in any clustering earthquake research yet. The results of the cluster of this method are compared with K-Means and K-Medoids method. That methods are validated by Davies-Boudin method (Davies & Bouldin, 1979). In particular, identify Gutenberg-Richer equation of Sulawesi earthquake of 1980 to early 2022, temporal and spatial analysis of *b*-value, efficiency and validation of clustering methods, temporal and spatial clustering, and spatial earthquake class are conducted. Python language program is implemented to perform this analysis.

Data and Methods

The earthquake data used in this research is secondary data obtained from the Incorporated Research Institution for Seismology (IRIS) by entering the coordinates or boundaries of the research area. IRIS data have been widely used in seismic analysis, namely earthquake risk (Rifa et al., 2020), regional tomography

earthquake (Koulakov et al., 2014), earthquake hazard (Papadopoulos & Kijko, 1991), etc. This data is used to analyze the characteristics and risks of the earthquakes that occurred over four decades in Sulawesi. During this decade, there are some destructive earthquakes, such as the earthquake in Donggala Regency in 2018, the earthquake in Mamuju Regency in 1984, and the earthquake in Mamuju Regency in 2021, etc. The earthquake events have been collected in the period 1980-2022, with the earthquake magnitude being more than equal to 5 SR, where there are 1039 data in that period. The data were analyzed using machine learning methods (Mini Batch K-Means and K-Medoids methods) and statistical method (maximum likelihood method). In machine learning, Mini Batch K-Means and K-Medoids methods are validated using the Davies Bouldin Index (DBI) method to determine the optimal number of clusters. The maximum Likelihood method is used to determine the pattern or level of activity of tectonic events (*b*-value and *a*-value).

The attributes used to cluster seismic events are earthquake epicenter, depth, and magnitude, while *a*-value and *b*-value analysis, earthquake epicenter, depth, magnitude, hours, minute, years, and month attributes are needed. To perform that analysis, the module of python language programming are pandas, geopandas, numpy, sci-kit-learn (Pedregosa et al., 2011), matplotlib, and seaborn. Meanwhile, temporal and spatial analysis of the *a*-value and *b*-value is conducted by the maximum likelihood method with zmap 6.0 free under Matlab language programming. The earthquake classifications are divided into six classes as shown in table 1. The earthquake maximum recorded by the seismometer in this data is magnitude 7.9 with serious damage effect as revealed in Table 1. The class of the earthquake in Table 1 are used to map the spatial distribution of the earthquake effect.

Table 1. Class and effect of the earthquake event.

Magnitude	Class	Effect
2.5 or less	Minor	Usually not felt but can be recorded by the seismograph
2.5 to 5.4	Light	Often felt, but only causes minor damage
5.5 to 6.0	Moderate	Slight damage to the building and other structures
6.1 to 6.9	Strong	May cause damage in a very populated area
7.0 to 7.9	Major	Serious damage
8.0 or greater	Great	Can totally destroy communities near the epicenter

Machine Learning Methods

Machine learning is a method for analyzing data mining, and is divided into supervised and unsupervised learning but in this paper, unsupervised learning, like Mini Batch K-Means and K-Medoids, is used to reach out the number clusters of an earthquake event. Before it is conducted, the number of clusters of data is determined by the elbow and silhouette method where the advantage of a

silhouette over than elbow is cluster structure and data distributions within the cluster as revealed in Table 2 (P. Rousseeuw et al., 1997).

Table 2. Interpretation of Silhouette coefficient.

Silhouette Coefficient (*SC*)	Interpretation
0.71 - 1.00	A strong structure
0.51 – 0.70	A reasonable structure
0.26 – 0.50	Weak structure
≤ 0.25	No substantial structure

Silhouette coefficient (SC) calculated by this formula (P. J. Rousseeuw, 1987):

$$\text{SC} = \frac{1}{n}\sum_{i=1}^{n} s(i) \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

where n is number of the earthquake event, $s(i)$ silhouette value (P. J. Rousseeuw, 1987). The numbers clusters in both elbow and silhouette methods have been validated by The Davies-Bouldin index (DBI) which is followed these formulas (Davies & Bouldin, 1979):

$$\text{DB} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n} \max_{j\neq i} \frac{\sigma_i + \sigma_j}{\tau_{i,j}} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

where DB is called the Davies-Bouldin index, d_i is the ratio of the qth root of the qth moment of the points in cluster i,j about the mean, $\sigma_i + \sigma_j$ with separation between cluster i,j, $\tau_{i,j}$.

Both Mini Bath K-Means and K-Medoids use the Euclidean distance to find centroid points within the earthquake event catalog. Mathematically the formula Euclidean distance follows equation 3.

$$d_{ij} = \sqrt{\sum_{a=1}^{p}\left(X_{ia} - X_{ja}\right)^2} \quad i=1,2,3,\dots,n; j=1,2,3,\dots,n \dots\dots\dots\dots\dots\dots\dots (3)$$

where d_{ij} is epicenter distance. The number iteration, epicenter coordinates attributes, $X_{ia} - X_{ja}$, depth attribute, and cluster number are used as input within Mini Batch K-Means Algorithm:

1: Given: k, mini-batch size b, iterations t, data set X
2: Initialize each c ∈ C with an x picked randomly from X
3: v ← 0
4: for i = 1 to t do
5: M ← b examples picked randomly from X
6: for x ∈ M do
7: d[x] ← f(C, x) // Cache the center nearest to x

8: end for
9: for x ∈ M do
10: c ← d[x] // Get cached center for this x
11: v[c] ← v[c] + 1 // Update per-center counts
12: η ← 1 v[c] // Get per-center learning rate
13: c ← (1 − η) c + ηx // Take gradient step;
14: end for;
15: end for

Mini Batch K-Means Method is more efficient in convergent than K-Means Method because the gradient descent method is applied to the optimization process on the updated centroid. Consequently, this method can be applied to big data analysis. to anticipate the outlier data of earthquake catalog, K-Medoids or partitioning around medoid method (PAM) is applied in this research. Like K-Means or Mini Batch K-Means methods, the K-Medoids method uses the same attributes as the input in its algorithm. In principle, the K-Medoids method divided n the earthquake data to be k cluster is a method that uses Euclidean distance to determine the medoid of each cluster. The algorithm of the K-Medoids method:

Step 1: Select the initial model
1: Given: K, iterations, data set X
2: Initialize each c ∈ C with an x picked randomly from X
3: v ← 0
4: for i = 1 to t do
5: M ← b examples picked randomly from X
6: for x ∈ M do
7: d[x] ← f(C, x) // Cache the center nearest to x
8: end for
9: Select k objects having the first k smallest values as initial medoids.
Step 2: (Update medoids)
10: 10:Find a new medoid of each cluster, which is the object minimizing the total distance to other
11: objects in its cluster. Update the current medoid in each cluster by replacing it with the new medoid.
Step 3: (Assign objects to medoids)
12: Assign each object to the nearest medoid and obtain the clustering result.
13: Calculate the sum of the distance from all objects to their medoids. If the sum is equal to the previous one, then stop the algorithm. Otherwise, go back to Step 1.

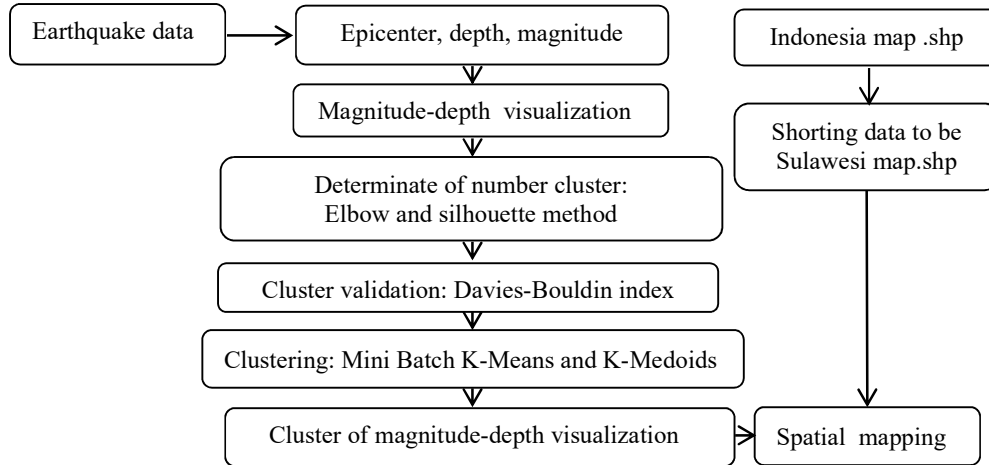A flowchart used to cluster earthquake event data and its spatial mapping is shown in Figure 1.

Figure 1. Flowchart of the clustering earthquake events.

The Maximum Likelihood Method and Estimated Damage

b-value called statistical parameters can describe the earthquake event which is caused by tectonic activities. The maximum likelihood method is used to calculate this value following Equation 4 (Kijko, 1988; Utsu, 1965):

$$b = \frac{\log e}{\overline{M}\text{-}M_{min}} = \frac{0.4343}{\overline{M}\text{-}M_{min}} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(4)$$

where $\overline{M}$ is the average of the magnitude of sample data, M_{min} is minimal of the magnitude in sample data. Then, the a-value is calculated by following the formula:

$$a = \log(N) + \log(b \ \ln 10) + M_{min}b \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \ (5)$$

b-value and *a*-value are the parameters that use to describe the seismicity caused by tectonic in the regional catalog of seismicity. The relations of these constants have been described by following linear formula (Gutenberg & Richter, 1944):

$$\log N(M) = a\text{-}bM \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(6)$$

where N is the number of events of magnitude M . In this research, calculated *a*-value and *b*-value by maximum likelihood or least square method are used to model Gutenberg-Richter construction, *b*-value temporal visualization, and spatial mapping of *a*-value and *b*-value. Furthermore to insight analysis of the earthquake effects, visualization estimated damage of these earthquakes have conducted in this research following the flowchart in Figure 2.
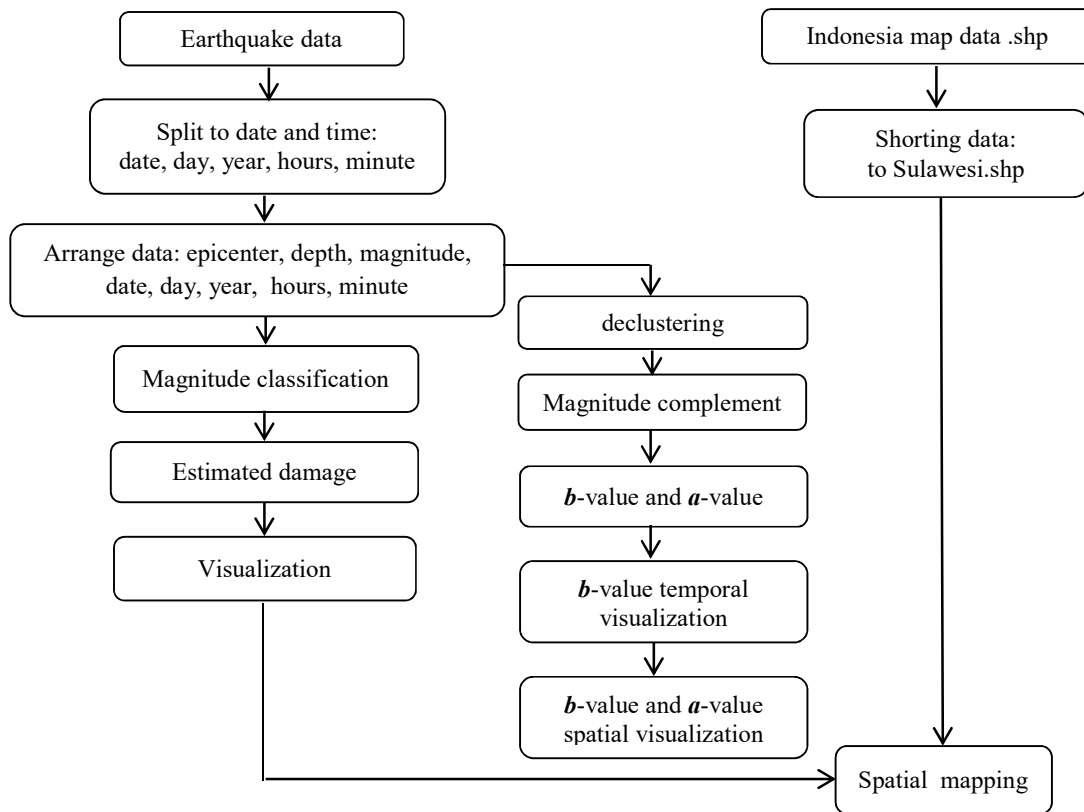
Figure 2. Flowchart of estimated damage and **b**-value and **a**-value analysis.

Results and Discussions
The Clustering Earthquake Events by Mini Batch K-Means and K-Medoids

The clustering of earthquake events has been successfully analysed or extracted information of the seismic events where the amount of the recorded data is 1039 events on magnitude ≥ 5 SR. The optimum number of the cluster has been obtained by Elbow and Silhouette coefficient for Mini Batch K-Means, K-Means, and K-Medoids as shown in Figure 3 and Figure 4. Python language programming is used to perform elbow and Silhouette coefficient analysis with the following script:

```
# Package of silhouette coefficient and davies Bouldin index
from sklearn.metrics import davies_bouldin_score
from sklearn.cluster import K-Means, Mini Batch K-Means
from sklearn_extra.cluster import KMedoids
from sklearn.metrics import silhouette_samples, silhouette_score
X = arr # variabel change to array
# Instantiate the clustering model, average score and visualizer
number_cluster = 3
model = KMedoids(number_cluster, random_state=42)
cluster_labels = model.fit_predict(X)
visualizer = SilhouetteVisualizer(model, colors='yellowbrick')
labels = model.labels_
```

```
visualizer.fit(X)        # Fit the data to the visualizer
visualizer.show()        # cluster visualization
# sihouette coefficent and davies Bouldin index
silhouette_avg = silhouette_score(X,cluster_labels)
val = davies_bouldin_score(X, cluster_labels )
print("The average silhouette_score is :", silhouette_avg )
print("Davies-Bouldin index (DBI) is :", val )
```
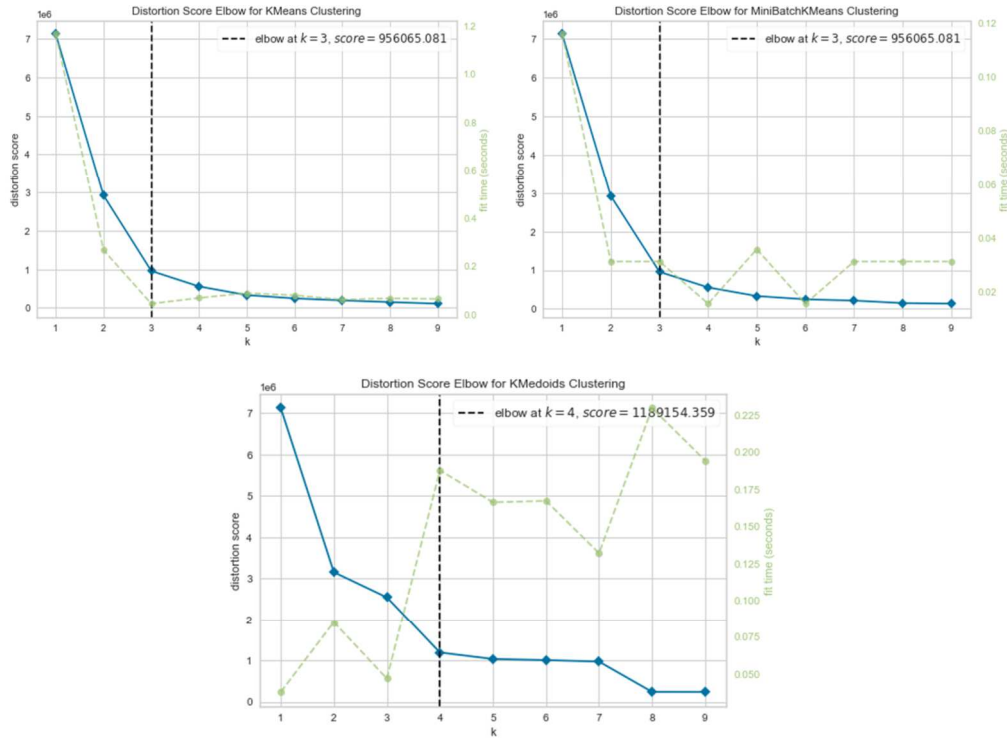

Figure 3. Elbow method using K-Means, Mini Batch K-Means, and K-Medoids algorithm on the earthquake data in the period of 1980 to 2022.

Based on Figure 3 and Table 3, The optimum number cluster of K-Means, Mini Batch K-Means, and K-Medoids are 3, 3, and 4 respectively, however, Mini Batch K-Means is more efficient in time or converges faster than the others because of applied gradient descent optimization on its algorithm (Pedregosa et al., 2011; Sculley, 2010). Gradient descent is one of the optimization algorithms used to find a local minimum of the earthquake attribute namely epicenter, magnitude, and depth.

Table 3. Optimum cluster and its time consumption.

Optimum cluster	Time consumption of clustering methods in second		
	K-Means	Mini Batch K-Means	K-Medoids
3	1.20	0.12	NaN
4	NaN	NaN	0.23

For more accuracy, the Silhouette method is applied to determine of the number clusters where the highest value of the Silhouette coefficient is used as the cluster number. Silhouette coefficient can be figured out by structure relations within the member clusters (Hidayati et al., 2021; Sculley, 2010). The other advantage of computing the Silhouette coefficient using python programming, the member size within each cluster can be interpreted descriptively as shown in Figure 4.

Table 4. Silhouette coefficient of K-Means, Mini Batch K-Means, and K-Medoids Method.

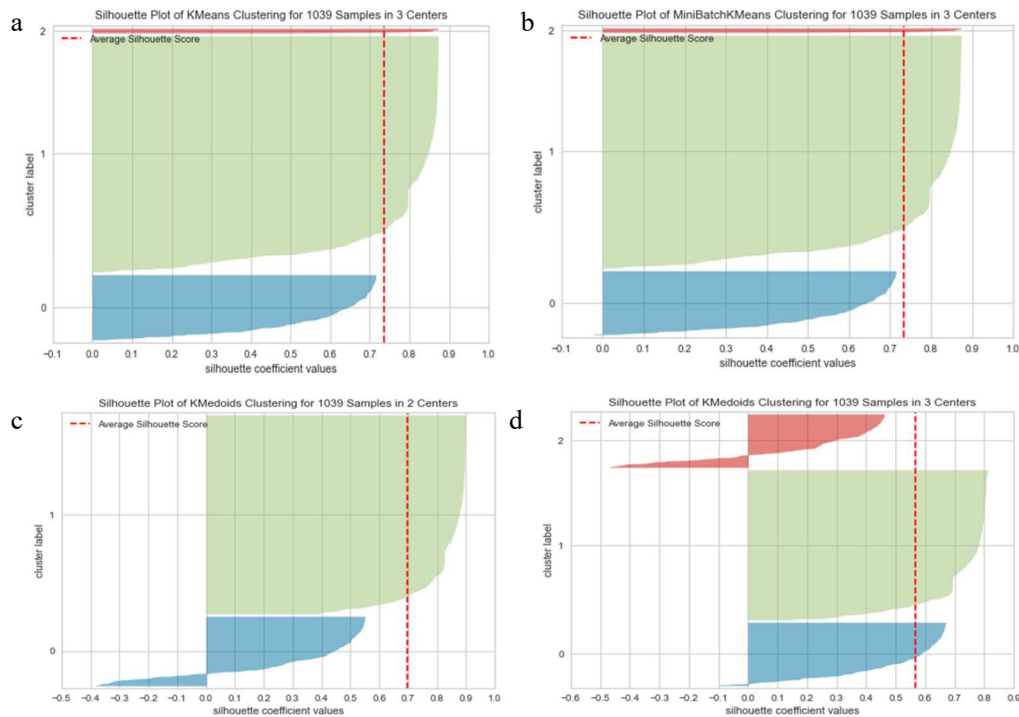| Cluster		Silhouette coefficient		
		K-Means	Mini Batch K-Means	K-Medoids
2		0.72	0.72	0.70
3		0.73	0.73	0.56
4		0.67	0.69	0.61
5	0.63	0.63	0.61	

Figure 4. Silhouette method earthquake data in the period of 1980 to 2022; (a) K-Means Method for 2 clusters, (b) Mini Batch K-Means for 3 clusters, (c) K-Medoids for 2 clusters, and (d) K-Medoids for 3 clusters.

Based on Table 4, the highest Silhouette coefficient on Both K-Means and Mini Batch K-Means is 0.73, and its member distribution is revealed in Figure 4. It means that earthquake events in each cluster have a strong similarity in the characteristic. Meanwhile K-Medoids method, the highest Silhouette coefficient is in 2 clusters with 0.70 Silhouette coefficient as in Table 4. Based on Table 2, this coefficient value is indicated that the cluster member has either a reasonable structure or quite a resemblance.

Based on the description of clustering the earthquake event, Both the Elbow and Silhouette coefficient method for Mini Batch K-Means and K-Means Method have a match in determining the optimum number of the cluster, namely 3 clusters. However, both the Elbow and Silhouette coefficient methods for K-Medoids are not in line to determine the optimum number cluster where the Elbow method is 4 clusters whereas the Silhouette coefficient is 4 clusters. In this case, Davies & Bouldin (1979) in a cluster separation measure had shown a method to validate of number clusters which are Davies-Bouldin Index (DBI). This method has been adopted by Ahmad & Irsalinda (2020) in Fuzzy K-Means with R programing language to validation of their models. In the same way, this research also applied the DBI technique to the number cluster validation for the K-Medoids method as in Table 5.

Table 5. Davies-Boudin index (DBI) of K-Means, Mini Batch K-Means, and K-Medoids method.

Cluster	DBI		
	K-Means	Mini Batch K-Means	K-Medoids
2	0.56	0.56	0.56
3	0.36	0.36	0.62
4	0.48	0.50	0.64
5	0.47	0.46	0.56

The lower value of DBI is indicating a better cluster (Davies & Bouldin, 1979; Pedregosa et al., 2011) so The optimum number of the cluster used to map temporal and spatial clusters are 3 clusters for Mini Batch K-Means and 2 clusters for K-Medoids.

In this research, both temporal and spatial mapping as in Figure 5 and Figure 6 have been conducted to analyze earthquake events in Sulawesi so that the earthquake information is more comprehensive. based on the depth, Fowler (2004) has stated that the earthquake event is divided into three classes namely shallow earthquakes with less than 70 km, intermediate earthquakes between 70 km up to 300 km, and deep earthquakes with more than 300 km.
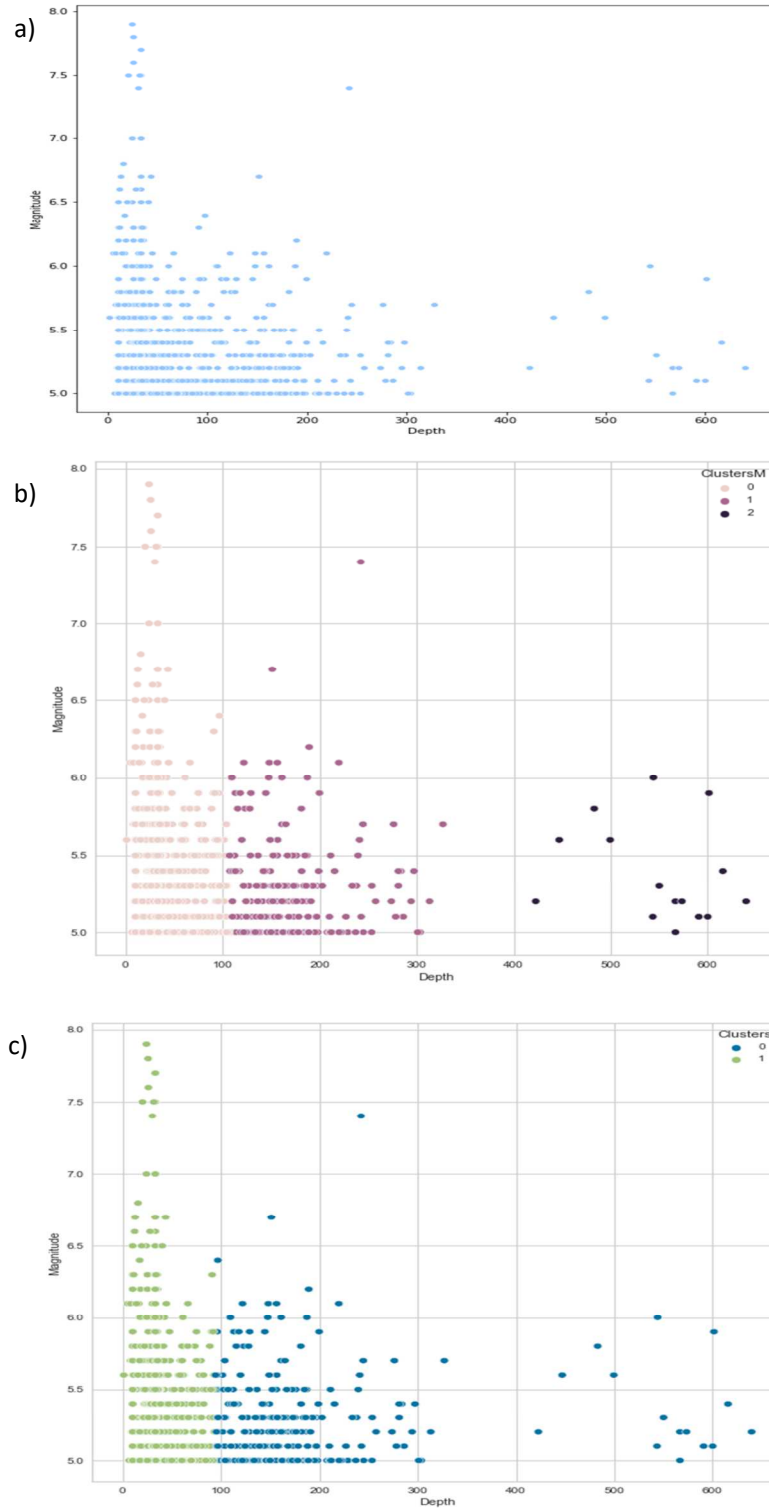
Figure 5. Temporal mapping of (a) uncluster, (b) 3 clusters of Mini Batch K-Means method, and (c) 2 clusters of K-Medoids method during the period 1980 - 2022.

Interesting results are depicted in Figures 5b and 5c where the pattern of the earthquake events are well segmented and have a similar classification as described by Fowler (2004), instead in Figure 5b. The moderate up to strong damages are in cluster 0 where its depth below 100 km, meanwhile the depth of the others cluster are in among above 100 km up to below 350 km (intermediate earthquakes with the average damage in moderate class), and above 350 (deep earthquake with the average damage is slight class. The K-Medoids method only recognized 2 classes, where cluster 0 below 100 km (shallow earthquake) and cluster 1 above 99 km. The members of cluster 1 are in the intermediate and deep earthquakes. The cluster 0 both Mini Bath K-Means and K-Medoids methods are the representation of the shallow earthquake where this cluster is more devastating than others because the earthquake event in this depth is close to the surface. Therefore the wave energy is received more destructive than others of the earthquake class. For example, Setiyono et al.(2019) recorded the shallow earthquake in Donggala-Palu-Sigi with 5.9 and 7.4 magnitudes of 10 km and 11 km in depth, respectively, are moderate and seriously destructive. Based on Figure 6a, cluster 0 are covered North Sulawesi, Gorontolo, Middle Sulawesi, and West Sulawesi, some regions of South Sulawesi, and Southeast Sulawesi Province, cluster 1 are covered North Sulawesi, Gorontolo, and some regions of Middle Sulawesi Province, and cluster 2 are covered some region of Gorontalo, Center Sulawesi, and Southeast Sulawesi Province.
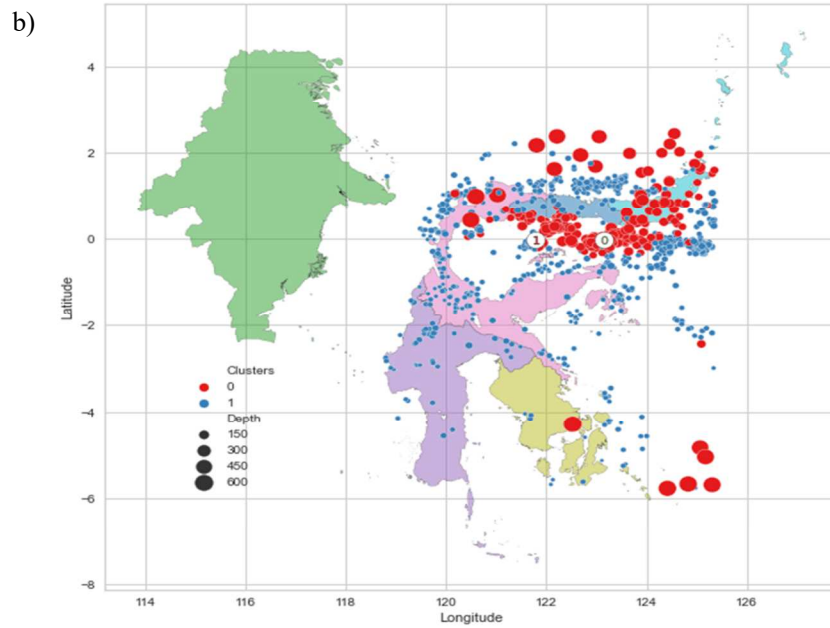
b)
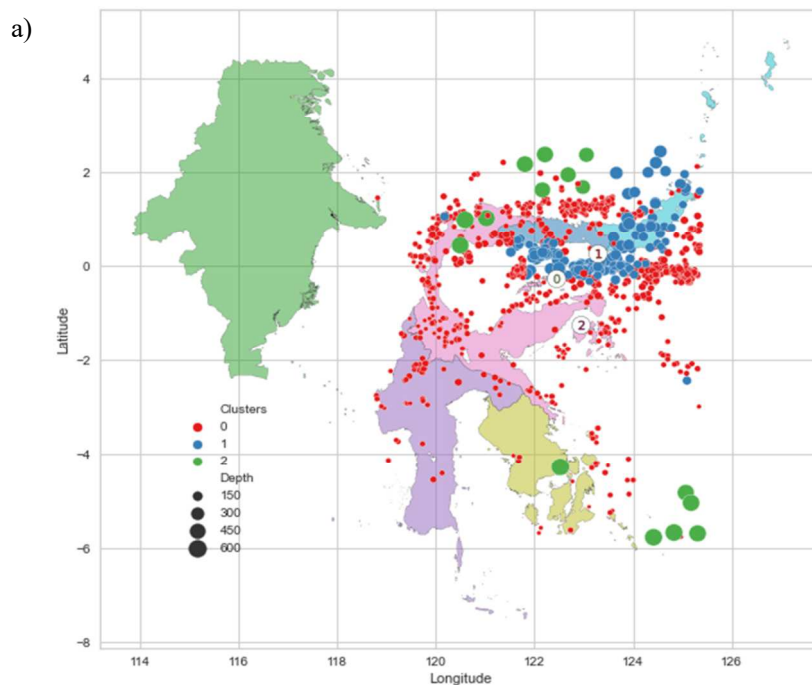
Figure 6. Spatial mapping of (a) cluster of Mini Batch K-Means method, and (b) clusters of K-Medoids method during the period 1980 - 2022.

Table 6. The centroid location of earthquake events during the period 1980 - 2022.

Centroid	Mini Batch K-Means			K-Medoids		
	Latitude	Longitude	Depth (km)	Latitude	Longitude	Depth (km)
0	-0.2655	122.4301	42.8592	-0.0250	121.7740	33.60
1	0.2655	123.2769	171.2994	-0.0240	123.1560	153.400
2	-1.2386	122.9326	552.9638			

The most dominant earthquake events occur in North Sulawesi, Gorontolo, Middle Sulawesi, and West Sulawesi Province, while the earthquake events in South Sulawesi Province rarely occurs in the others so that disaster caused by the earthquake probably minimum happened in this province. Based on K-Medoids in either Figure 5c or Figure 6b, the effect of earthquake caused by cluster 0 is not felt in West and South Sulawesi because of the distance of the earthquake events (extremely attenuated wave) and the earthquake type (intermediate earthquake). This result has a similarity in cluster 1 with Mini Batch K-Means Method. The centroid point of each cluster is revealed in Table 6, where the shallow earthquake is represented by centroid 0, the intermediate earthquake is represented by centroid 1, and the deep earthquake is represented by centroid 2. The seismicity activities of the earthquake of each cluster are described by *a*-value and *b*-value parameters.

The *a* and *b*-Value and Estimated Damage

Cluster 0 in this paper is hight destructive than the others because cluster members are in the shallow earthquake. During the period 1980-2022, the percentage of moderate, major, and strong earthquake events was recorded above 80 percent, under 10 percent, and under 5 percent, respectively, as shown in Figure 7a. Other than that, slight, severe, and serious damage s was recorded above 80 percent, under 10 percent, and under 5 percent, respectively as shown in Figure 7b. For example, the Slight and strong damage earthquake in Manado with magnitude 5.3 SR at 26 km, magnitude 5.2 SR at 33 km, and magnitude 6.7 SR at 63 km was confirmed by Setiyono et al.(2019). The damage caused by the earthquakes, namely slight damage to some buildings, died, injured, and bridges. Another example of serious damage with magnitude 7.4 SR at 11 km was recorded in Donggala-Palu-Sigi which damage caused by this disaster were deaths, buried, injured and lost people, damaged buildings, etc. The frequency of earthquake events and damage classification as shown in Figures 7c and 7d, respectively.
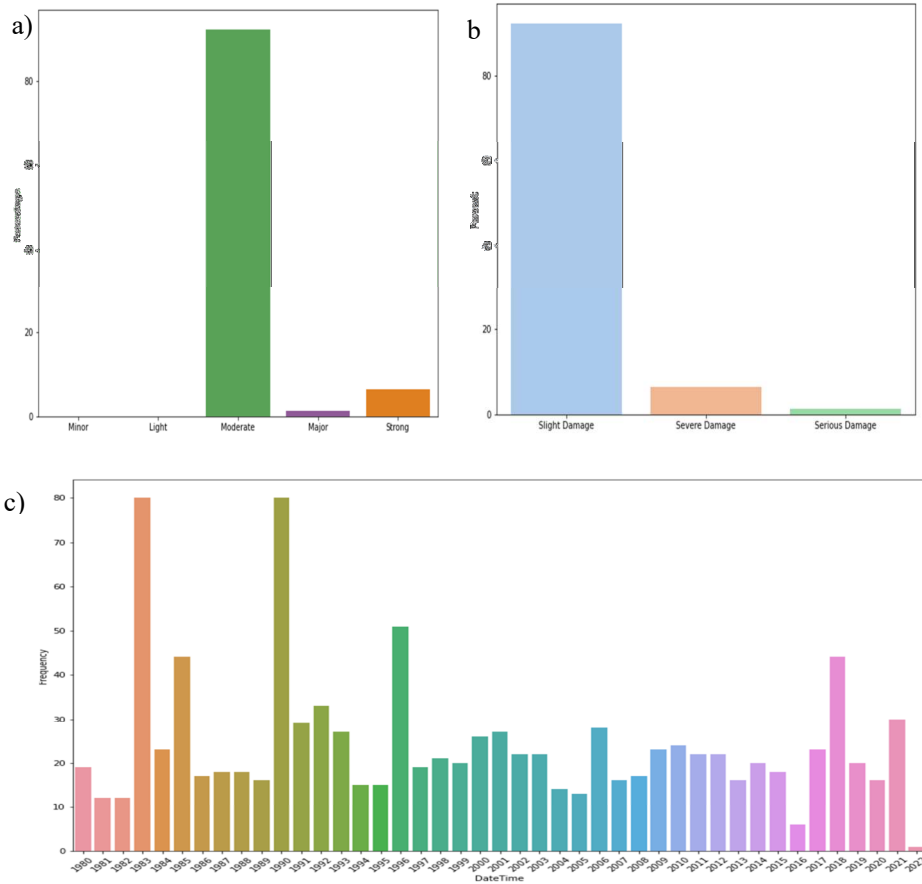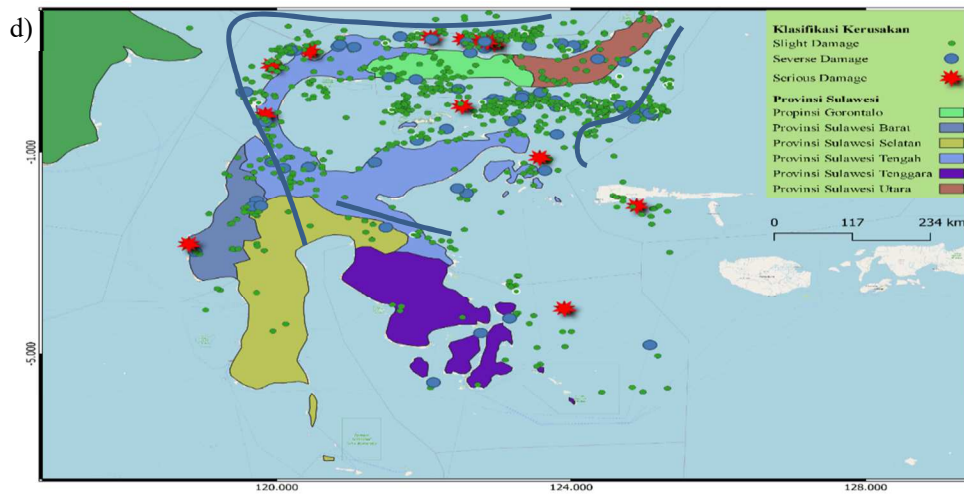
Figure 7. The temporal and spatial mapping of (a) earthquake categories, (b) damage percentage, (c) the earthquake frequency, and (d) damaged classifications.

Based on Figure 7c and 7d, most of the destructive earthquake event has occurred in the Palu-Koro fault and North Sulawesi Trench. Palu-Koro fault as the source of the earthquake on land is a major active fault in Sulawesi whereas North Sulawesi Trench is the source of the earthquake offshore (Supartoyo et al., 2014). Besides major faults, local activate faults in West Sulawesi are disasters source caused by shallow earthquakes that occurred periodically in 1980 and 2021.
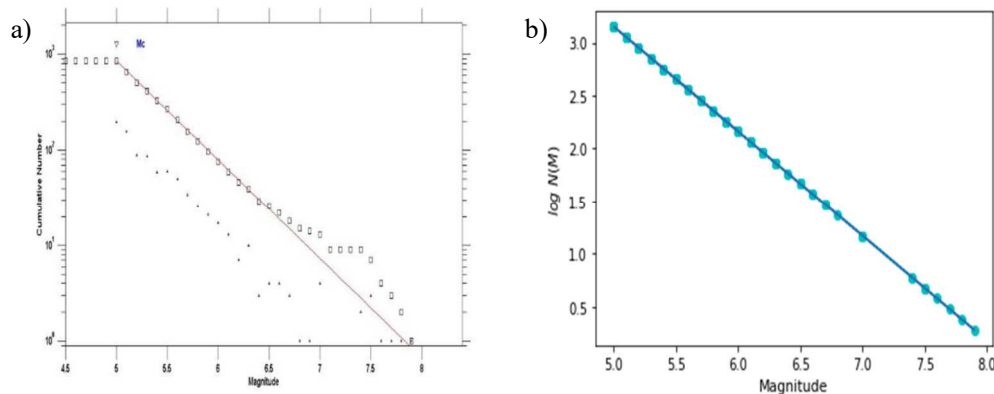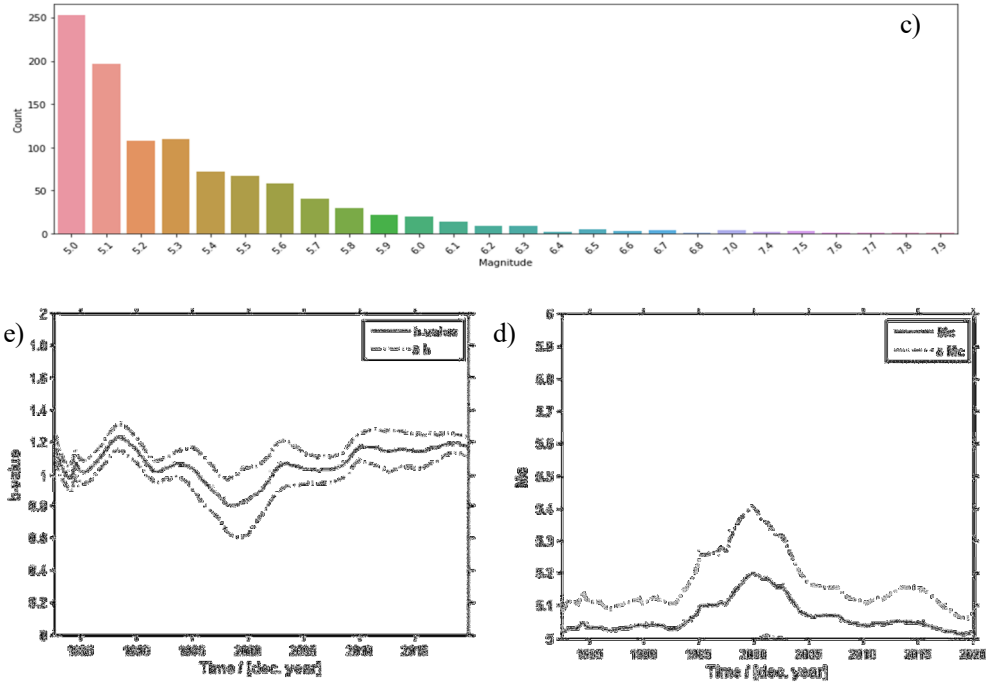
Figure 8. The Temporal variations of (a) the cumulative number of the magnitude, (b) the Gutenberg-Richer model, (c) the Magnitude number, (d) magnitude complement, and (e) *b*-value temporal.

The effects of the earthquake events have been revealed in Figure 8, where the earthquake events are relatively related to tectonic activities. The frequency-magnitude relation caused by tectonic activity expressed by Gutenberg & Richer (1944) is reciprocal as shown in Figures 8a, 8b and 8c. The magnitude complement (Mc) of the earthquake event is 5 SR. This value is similar to the magnitude minimum in Figure 8b, so the magnitude complement is the lowest magnitude or minimum frequency of the earthquake events. The temporal magnitude complement (Mc) variations in Figure 8d showed the minimum wave energy on the subduction zone is diverse (Scholz, 2015), and the minimum wave energy increased pattern in the period 1990 to 2005 shows the energy gradient of the earthquake is relatively both strong and destructive. From the Mc value, the *a*-value was obtained at 8.1 that is indicated strong seismotectonic activities in the subduction zones, and the *b*-value obtained at 0.99 indicated moderate level stress on the subduction zones. So, equation 6 is written to:

$$\log N(M) = 81\text{-}0.99M \dots \dots \dots (7)$$

Scholz (1968) stated that *b*-values are reciprocal with the accumulated stress and *a*-values are linear with seismicity activities. Other than that, Scholz (1968) also stated that volcanic origins are characterized by *b*-value between 0

.

and 1. Based on this statement, the earthquake events in this research are mostly on plutonic and volcanic rocks (White et al., 2017). There are a major minimum curve and two minor minimum curves as shown in Figure 8e. The first minimum that occurred during the period 1980-1989 was increasing either force or accumulated stress and decreasing either force or release energy (Rigo et al., 2018; Scholz, 2015). In the major earthquakes of 1984 and 1986, there was a decreasing gradient of the b-value. This pattern was repeated during the period 1990-2001, so the energy of the seismic wave was released. The increasing b-value gradient that happened during the period 2001-2007 was stress accumulations. This stress was released in 2008 where it is characterized by a decrease in the b-value gradient. Based on this description, it might occur in any year periodically. Both the b-value and a-value are spatially shown in Figure 9 which this mapping is depicted the b-value and a-value distributions of each province in Sulawesi.
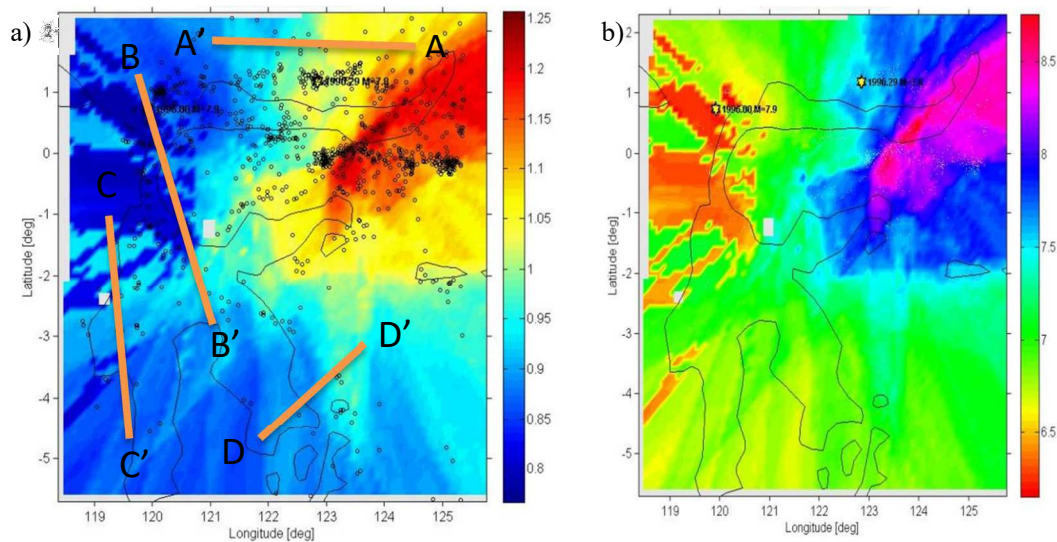

Figure 9. Spatial variation of (a) b-value, and (b) a-value during the period 1980-2022.

For b-value interpretation in Figure 9, there are used four cross-sections namely North Sulawesi-Gorontalo-Center Sulawesi (cross-section $A - A'$) high decreasing gradient b-value, Middle Sulawesi to Southeast Sulawesi (cross-section $B - B'$) decreasing and increasing b-value, Middle Sulawesi-West Sulawesi-South Sulawesi (cross-section $C - C'$) increasing-decreasing-increasing b-value, Southeast Sulawesi (cross-section $D - D'$) decreasing b-value. North Sulawesi-Gorontalo-Middle Sulawesi (cross section $A - A'$) have released wave energy because of seismic activities on North Sulawesi Trench. Cross section $B - B'$ is high-stress accumulation on the subduction zones so this Middle Sulawesi have the potential to release high energy of the seismic waves, and have

the level of rock fragility (Scholz, 1968). Seismic activities in this province are caused by active fault Palu-Koro and local faults (Supartoyo et al., 2014). The stress accumulations are also in some regions of West Sulawesi caused by normal fault on the fold-thrust-belt with quaternary sedimentary rock, and genozoic volcanic and plutonic rock (White et al., 2017; Yanuarto, 2021). South Sulawesi is relatively uniform in both *b*-value and a-value distribution, so earthquake events in this are relatively rare than in the others province. In this province, the highest *b*-value,1, caused by the active Mentano fault have lower seismic activities as shown in Figure 9b (Kurniawati & Rahman, 2020). It is an inherently seismic activity in several areas in South Sulawesi. Furthermore, The earthquake activity pattern in Southeast Sulawesi is the same as in South Sulawesi.

Earthquake event analysis has conducted by Mini Batch K-Means, K-Medoids, and maximum likelihood method where its result showed that the disasters caused by the earthquake event in the province of Sulawesi are dominant in cluster 0 (shallow earthquake) with a *b*-value below 0.9 and *a*-value below 7.5. The disaster caused by the earthquake event in the province of Sulawesi is dominant in cluster 0 (shallow earthquake) with *b*-value below 0.9 and *a*-value below 0.75. The provinces in Sulawesi with this value and cluster have indicated that its region has high accumulated energy and makes a disaster, respectively. The disaster in Sulawesi by three methods is shown in Table 7.

Table 7. Risk analysis of the earthquake in Sulawesi based on three methods.

Province	Mini Batch K-Means Clusters			K-Medoids Clusters		Maximum Likelihood (*b*-value)
	0	1	2	0	1	
North Sulawesi	√	√√	x	√	√√	√
Gorontalo	√	√√	x	√	√√	√
Middle Sulawesi	√	*x*	x	√	*x*	√
Southeast Sulawesi	√√	x	x	√√	*x*	√√
West Sulawesi	√	x	x	√	*x*	√
South Sulawesi	√√	x	x	√√	*x*	√√

√ : Severe up to serious damage and the high seismicity of the earthquake
√√ : Slight up to severe damage and moderate seismicity of the earthquake
x : Minor effect of the earthquake

Based on Mini Batch K-Means in cluster 0, K-Medoids in cluster 0, and the maximum likelihood method in Table 7, North Sulawesi, Gorontalo, Middle Sulawesi, and West Sulawesi Province are classified into severe damage by the earthquake events. While Southeast Sulawesi and South Sulawesi provinces are classified into slight up to severe damage. The earthquake events in cluster 1 both

Mini Batch K-Medoids is classified into minor effect up to slight although these regions have high seismicity of the earthquake.

Conclusion

The Analysis pattern of the earthquake events is successfully conducted by Mini Batch K-Means and K-Medoids method (machine learning), and the maximum likelihood method. Other than that, Mini Batch K-Means and K-Medoids method have been validated by Davies-Bouldin Index. While the maximum likelihood method is used to calculate the a-value and b-value. Mini Batch K-Means has classified the depth of the earthquake events into shallow, intermediate, and deep earthquakes, while the K-Medoids method can classifications of the earthquake events into shallow earthquakes and earthquake events above 100 km. Based on the Davies-Bouldin index and silhouette coefficient, the Mini Batch K-Means method is more robust than K-Medoids Method, instead, the earthquake events in Sulawesi with magnitude ≥ 5 during the period 1980 - early 2022. Furthermore, the classifications of the earthquake event based on Mini Batch K-Means are more relevant to Fowler's concept (2004), and time efficient than both the K-Means and K-Medoids Methods.

Both Mini Batch K-Means and K-Medoids methods have depicted the pattern of earthquake events in Sulawesi, where the earthquake events are categorized as dangerous natural hazards because both serious and major earthquakes are shallow earthquake clusters. The Seismicities and stress level on the subduction zones is figured out by a-value and b-value both temporal and spatial. The b-value and a-value obtained in Sulawesi are 0.99 and 8.1, repeatedly. These values are shown that stress and seismicities are so enough strong. Based on the b-value temporal, the earthquake events are periodic events, where during the periods 1990-2001 was major release energy and during the periods 2001-2007 was major accumulated stress on subduction zones. The regions with a b-value and a-value less than 0.9 and 7.5, respectively, and in cluster 0, namely the western part of North Sulawesi, Gorontalo, Middle Sulawesi, and West Sulawesi Province, are as vulnerable to earthquake disasters. Meanwhile, the region in cluster 1 and cluster 2 with a b-value and a-value more than 0.9 and 7.5 respectively namely South Sulawesi, the Northern part of North Sulawesi, and Southeast Sulawesi Province, are categorized as minor earthquake disasters.

The results of this study can provide an overview of seismic activities and their impact in Sulawesi, but a comprehensive analysis focal mechanism analysis to determine the fault and slip of structural geology can be conducted. This analysis is using waveform seismic data. For clustering with machine learning, Mini Batch K-Means clustering can be applied to cluster the seismic event, and this method can be used to cluster big data because more efficient than the K-Means, and K-Medoids Methods. Deep and machine learning can be combined for pattern prediction of seismic events.

Acknowledgements

References

Ahmad, A., & Irsalinda, N. (2020). Cluster Analysis of Earthquake's Data Clustering in Indonesia using Fuzzy K-means Clustering. *Proceeding International Conference on Science and Engineering*, *3*, 3–7. https://doi.org/10.14421/icse.v3.457

Cipta, A., Robiana, R., Griffin, J., Horspool, N., Hidayati, S., & Cummins, P. (2016). A probabilistic seismic hazard assessment for Sulawesi, Indonesia. *Geological Society, London, Special Publications*, *441*. https://doi.org/10.1144/SP441.6

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Ernandi, F. N., & Madlazim. (2020). Analisis Variasi a-Value dan b-Value dengan Menggunakan Software ZMAP V.6 sebagai Indikator Potensi Gempa Bumi di Wilayah Nusa Tenggara Barat. *Inovasi Fisika Indonesia*, *9*(3), 24–30. https://doi.org/10.26740/ifi.v9n3.p24-30

Eva, M. N. (2016). Pemetaan Daerah Rentan Gempa Bumi sebagai Dasar Perencanaan Tata Ruang dan Wilayah di Provinsi Sulawesi Barat. *KURVATEK*, *1*(2), 41–47. https://doi.org/10.33579/krvtk.v1i2.417

Fowler, C. M. R. (2004). *The Solid Earth: An Introduction to Global Geophysics* (2nd edition). Cambridge University Press.

Gutenberg, B., & Richter, C. F. (1944). Frequency of earthquakes in California*. *Bulletin of the Seismological Society of America*, *34*(4), 185–188. https://doi.org/10.1785/BSSA0340040185

Hidayati, R., Zubair, A., Pratama, A. H., & Indana, L. (2021). Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering. *Techno.Com*, *20*(2), 186–197. https://doi.org/10.33633/tc.v20i2.4556

Koulakov, I., Maksotova, G., Mukhopadhyay, S., Raoof, J., Kayal, J., Jakovlev, A. V., & Василевский, A. (2014). Variations of the crustal thickness in Nepal Himalayas based on tomographic inversion of regional earthquake data. *Solid Earth Discussions*, *6*. https://doi.org/10.5194/sed-6-2867-2014

Kijko, A. (1988). Maximum likelihood estimation of Gutenberg-Richterb parameter for uncertain magnitude values. *Pure and Applied Geophysics*, *127*(4), 573–579. https://doi.org/10.1007/BF00881745

Kurniawati, I., & Muhammad Rahman, R. (2020). Analisis Statistik Kegempaan Wilayah Sulawesi Selatan Periode 2006-2018. *Jurnal MSA (Matematika*

Elkawnie: Journal of Islamic Science and Technology Vol. 9, No. 1, June 2023
(www.jurnal.ar-raniry.ac.id/index.php/elkawnie)

DOI: 10.22373/ekw.v9i1.13027 | 21

Dan Statistika Serta Aplikasinya), *8*(1), 77–81. https://doi.org/10.24252/msa.v8i1.12942

Kurniawati, I., & Rahman, M. (2020). Analisis Statistik Kegempaan Wilayah Sulawesi Selatan Periode 2006-2018. *Jurnal MSA (Matematika Dan Statistika Serta Aplikasinya)*, *8*, 77. https://doi.org/10.24252/msa.v8i1.12942

Novinanti, P., Setyorini, D., & Rafflesia, U. (2017). K-Means cluster analysis in earthquake epicenter clustering. *International Journal of Advances in Intelligent Informatics*, *3*, 81. https://doi.org/10.26555/ijain.v3i2.100

Nurana, I., Simanjuntak, A. V. H., Umar, M., Kuncoro, D. C., Syamsidik, S., & Asnawi, Y. (2021). Spatial Temporal Condition of Recent Seismicity In The Northern Part of Sumatra. *Elkawnie: Journal of Islamic Science and Technology*, *7*(1), 131–145. https://doi.org/10.22373/ekw.v7i1.8797

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Ray, S., Alam, Md. J. B., Haque, M., Das, S. K., Tanmoy, B. B., & Hasan, Md. N. (2019). A study on b-value and investigation of seismic hazard in Sylhet seismic region, Bangladesh using Gumbel's extreme value distribution method. *SN Applied Sciences*, *1*(5), 435. https://doi.org/10.1007/s42452-019-0442-3

Rifa, I. H., Pratiwi, H., & Respatiwulan, R. (2020). Clustering of Earthquake Risk in Indonesia using K-Medoid dan K-Mean Alghorithms. *MEDIA STATISTIKA*, *13*(2), 194–205. https://doi.org/10.14710/medstat.13.2.194-205

Rigo, A., Souriau, A., & Sylvander, M. (2018). Spatial variations of b-value and crustal stress in the Pyrenees. *Journal of Seismology*, *22*, 337–352. https://doi.org/10.1007/s10950-017-9709-6

Rousseeuw, P., Hubert, M., & Struyf, A. (1997). Clustering in an Object-Oriented Environment. *Journal of Statistical Software*, *01*. https://doi.org/10.18637/jss.v001.i04

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

Scholz, C. (1968). *The Frequency-Magnitude Relation of Microfracturing in Rock and its Relation to Earthquakes*. *58*. https://doi.org/10.1785/BSSA0580010399

Scholz, C. (2015). On the Stress Dependence of the Earthquake b-value. *Geophysical Research Letters*, *42*. https://doi.org/10.1002/2014GL062863

Sculley, D. (2010). Web-scale k-means clustering. *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 1177. https://doi.org/10.1145/1772690.1772862

Setiyono, U., Gunawan, I., Yatimantoro, T., Imananta, R. T., Ramadhan, M., Hidayanti, Angraini, S., Hawati, P., & Yogaswara, D. S. (2019). *Katalog gempabumi signifikan dan merusak 1821—2018* (Jakarta; Vol. 1). Badan Meteorologi Klimatologi dan Geofisika.

Seydoux, L., Balestriero, R., Poli, P., Hoop, M. de, Campillo, M., & Baraniuk, R. (2020). Clustering earthquake signals and background noises in continuous seismic data with unsupervised deep learning. *Nature Communications*, *11*(1), 3972. https://doi.org/10.1038/s41467-020-17841-x

Supartoyo, S., Sulaiman, C., & Junaedi, D. (2014). Kelas tektonik sesar Palu Koro, Sulawesi Tengah. *Jurnal Lingkungan Dan Bencana Geologi*, *5*(2), 111–128. https://doi.org/10.34126/jlbg.v5i2.68

Utsu, T. (1965). A method for determining the value of b in a formula log n=a-bM showing the magnitude frequency relation for earthquakes. *Geophys. Bull. Hokkaido Univ*, *13*, 99–103.

White, L. T., Hall, R., Armstrong, R. A., Barber, A. J., BouDagher Fadel, M., Baxter, A., Wakita, K., Manning, C., & Soesilo, J. (2017). The geological history of the Latimojong region of western Sulawesi, Indonesia. *Journal of Asian Earth Sciences*, *138*, 72–91. https://doi.org/10.1016/j.jseaes.2017.02.005

Yanuarto, T. (2021). *Pembelajaran Gempa Sulbar, Fenonema dan Dampak Kerusakan Bangunan* [BNPB]. https://bnpb.go.id/berita/pembelajaran-gempa-sulbar-fenonema-dan-dampak-kerusakan-bangunan

Yuan, R. (2021). An improved K-means clustering algorithm for global earthquake catalogs and earthquake magnitude prediction. *Journal of Seismology*, *25*(3). https://doi.org/10.1007/s10950-021-09999-8

Zakaria, Z., & Sidarto, S. (2015). Aktifitas Tektonik di Sulawesi dan Sekitarnya Sejak Mesozoikum Hingga Kini Sebagai Akibat Interaksi Aktifitas Tektonik Lempeng Tektonik Utama di Sekitarnya. *Jurnal Geologi dan Sumberdaya Mineral*, *16*(3), 115–127. https://doi.org/10.33332/jgsm.geologi.v16i3.36