Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 4 (2024) 454 - 465 e-ISSN: 2580-0760

Analysis of Sulawesi Earthquake Data from 2019 to 2023 using DBSCAN Clustering

Ody Octora Wijaya[1*], Rushendra[2]
[1, 2] Informatics, Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia
[1]odyoctorawijaya@gmail.com, [2]rushendra@mercubuana.ac.id

Abstract

Sulawesi is a region in Indonesia known for its significant seismic activity, and its history of impactful earthquakes makes it *an area of crucial importance for in-depth analysis. This study analyses earthquake occurrence data in the Sulawesi region from 2019 to 2023 using clustering methods with the DBSCAN algorithm. The utilization of the DBSCAN algorithm was chosen for its ability to cluster data based on spatial density, well-suited for analyzing the spatial patterns of earthquakes. DBSCAN is known for its effectiveness in identifying spatial clusters, especially in handling data with undefined density patterns. The primary aim of this research is to identify spatial earthquake occurrence patterns, classify regions with similar earthquake occurrence rates, describe the characteristics of the resulting spatial clusters, and identify seismic gap areas. The results of analysis and clustering using the DBSCAN algorithm have identified clusters with earthquake depth characteristics, which are expected to make a significant contribution to mapping and understanding earthquake vulnerability and distribution in this region. These findings can aid in more effective disaster mitigation planning, support sustainable development efforts, and enhance earthquake preparedness and response in Sulawesi. This study contributes to a better understanding of earthquake patterns and potential seismic gaps in Sulawesi, which is crucial for developing improved risk mitigation strategies and supporting sustainable development policies.*

Keywords: clustering; DBSCAN; earthquake; Sulawesi; seismic gap

How to Cite: O. O. Wijaya and Rushendra, "Analysis of Sulawesi Earthquake Data from 2019 to 2023 using DBSCAN Clustering", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 4, pp. 454 - 465, Aug. 2024.
DOI: https://doi.org/10.29207/resti.v8i4.5819

1. Introduction

The Sulawesi region in Indonesia has a significant seismic history, with a series of earthquakes that can have serious impacts on life and infrastructure. Studying earthquake history over some time is crucial for understanding earthquake patterns, assessing potential risks, and contributing to disaster mitigation efforts [1].

Sulawesi Island is one of the islands located in Indonesia and is part of the Pacific Ring of Fire. This island has high geological complexity and is considered one of the regions with significant tectonic fault activity in Indonesia [2]. Research on active tectonic faults in Sulawesi Island is important for understanding and mitigating the risk of earthquake-related disasters and potential tsunamis. Some of the active faults that serve as earthquake source zones include the Palu-Koro Fault and Waianae Fault in western Sulawesi. Additionally, the Matano Fault and Lawanopo Fault are in eastern Sulawesi, and the Gorontalo Fault is in northern Sulawesi [3].

This research is a crucial step in evaluating seismic activity patterns in the Sulawesi region. By utilizing clustering techniques, we aim to identify spatial and temporal patterns of earthquake events. These patterns may indicate areas with low activity or seismic gaps within densely seismic regions. Through a deeper understanding of the earthquake activity patterns discovered, this study can provide valuable insights into the potential existence of seismic gap areas in Sulawesi.

Previous research has also addressed earthquake datasets from the period 2018 to 2020, obtained from the BMKG repository website for the entire Indonesian region. Using the DBSCAN algorithm, this research formed four clusters with four noise points and achieved a Silhouette Coefficient Score of 0.35 [4].

Several previous studies have demonstrated that using the DBSCAN algorithm can yield a Silhouette Score of

Received: 24-05-2024 | Accepted: 27-07-2024 | Published Online: 04-08-2024

0.81091 and a Gamma Index of 0.98104, indicating the reliability of the DBSCAN algorithm in processing earthquake data. However, this study processed the dataset for the entire Indonesian region in 2020. Other studies have also processed datasets sourced from the USGS for the Indonesian region from 2012 to 2021. By applying various filters based on parameters such as magnitude value and depth, these studies resulted in a Silhouette Score of 0.73 [5].

Several studies have also described earthquake data processing using the DBSCAN clustering method in the West Java region in 2021, achieving a Silhouette Coefficient Score of 0.713 [6].

In the northern Sulawesi region, earthquake data analysis has been conducted using agglomerative clustering methods [7].

In the Bali region, the DBSCAN method has been utilized to process earthquake data and generate mappings of earthquake potential zones with a fairly good cluster validity level [8].

Not only in Indonesia but earthquake data analysis in Greece has also been conducted in previous research using a new clustering method called MAP-DBSCAN. This method, which refers to the DBSCAN method, was used to group seismic zones in Greece during the period from 2012 to 2019 [9].

Previous studies have highlighted the reliability of the DBSCAN algorithm in earthquake data analysis, with several research noting high Silhouette Score and Gamma Index values. However, the focus of these studies often extends beyond the Sulawesi region, such as research utilizing the 2020 dataset for the entire Indonesian region. Additionally, although there are studies involving Indonesia, including West Java, that employ the DBSCAN algorithm, their analysis tends to be limited to shorter periods. Hence, there is a gap in the adequately researched analysis of earthquake data in the Sulawesi region from 2019 to 2023 using clustering methods with the DBSCAN algorithm.

This research aims to investigate seismic patterns in the Sulawesi region over the past four years using clustering methods with the DBSCAN algorithm. The main objectives of this study are to identify crustal movement patterns, pinpoint areas vulnerable to earthquakes (seismic gaps), and measure seismicity levels in the area [10].

Additionally, the research aims to apply clustering methods with the DBSCAN algorithm to categorize earthquake data and provide a better understanding of earthquake-prone areas in Sulawesi. Consequently, this study is expected to significantly contribute to earthquake risk mitigation efforts and disaster preparedness strategy development in the Sulawesi region [11].

2. Research Methods

The Meteorology, Climatology and Geophysics Agency (BMKG) is the authorized institution in Indonesia for monitoring earthquake activity. We have also had discussions with several colleagues at BMKG regarding this research.

For the workflow of this research, the overview is provided in the form of a flowchart. This is to facilitate researchers in carrying out the research stages. A research workflow is illustrated in Figure 1.

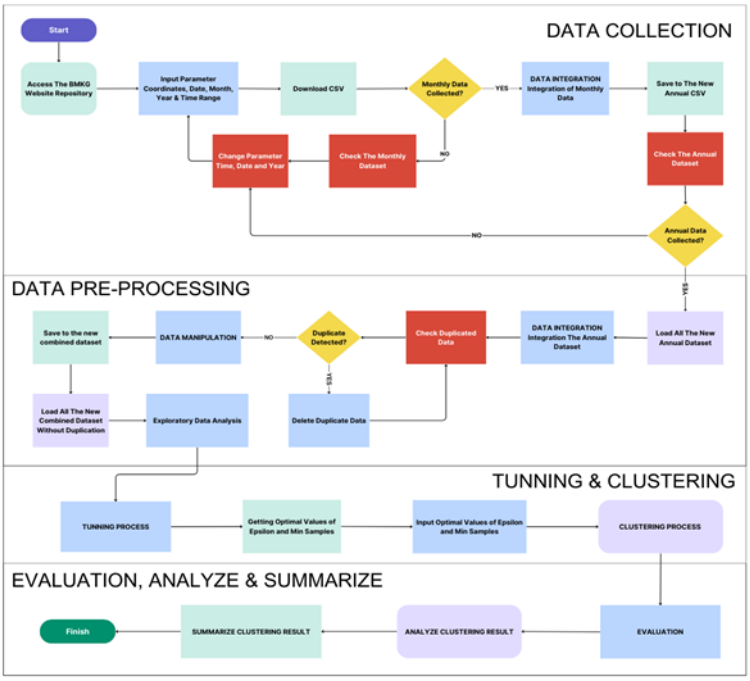

Figure 1. Research Workflow

2.1 Data collection

The first step is data collection. The dataset for this study consists of earthquake occurrences in the Sulawesi region from January 2019 to December 2023, obtained from the public earthquake repository website owned by BMKG, totalling 10,255 records.

The earthquake dataset from the BMKG repository website has 12 columns including: (1). No, (2). Event ID, (3). Date time, (4). Latitude, (5). Longitude, (6). Magnitude, (7). Mag Type (8). Depth (km), (9). Phase Count, (10). Azimuth Gap, (11). Location, (12). Agency.

2.2 Data pre-processing

The second step is data pre-processing, which is carried out to merge datasets, identify and handle anomalies in the collected dataset [12]. Activities in this step include data cleaning, where the process removes duplicate data based on the "Event ID" column in the dataset.

Data transformation is also performed in this data pre-processing step, which involves renumbering the “No” column of the collected and merged dataset.

2.3 Exploratory data analysis

The third step is Exploratory Data Analysis (EDA), conducted to identify patterns in the dataset, examine the statistics of the dataset after performing data pre-processing [13], and form initial hypotheses that can be tested further. Table 1 shows the statistics of the collected dataset after performing data pre-processing.

EDA also involves visualizing data through graphs and plots to facilitate understanding and analysis. By performing EDA, we can gain critical preliminary insights before conducting more complex statistical analysis or modelling [14].

Table 1. Shows the statistics of the collected dataset after performing data pre-processing.

A descriptive analysis of 10,238 earthquake events indicates that the average magnitude is 121.85041, with relatively small variation. Earthquake depths range from 0.813944 km to 6.778839 km, with an average depth of 3.237469 km. The data distribution shows that most earthquakes occur at shallow depths, providing crucial insights for further seismic analysis.

Table 1. Statistic of dataset after data pre-processing

	No	Latitude	Longitude	Magnitude	Depth (km)	Phase Count	Azimuth Gap
Count	10238.000	10238.00	10238.00	10238.00	10238.00	10238.00	10238.00
Mean	5119.5000	1.099820	121.850410	3.237469	38.328482	24.492772	117.350120
STD	2955.6000	1.582566	1.668430	0.747337	53.154834	28.628435	57.348734
Min	1.0000	-7.532882	117.953980	0.813944	1.000000	4.000000	8.606781
25%	2560.2500	-2.253999	120.438528	2.707554	10.000000	11.000000	72.962211
50%	5119.5000	-0.950472	121.942909	3.132503	10.000000	17.000000	103.618873
75%	7678.7500	0.052527	122.940958	3.665022	37.000000	28.000000	151.649754
Max	10238.0000	2.907121	125.857842	6.778839	298.00000	465.00000	348.699799

In the exploratory data analysis process, we obtained information about the distribution of earthquake occurrences each month during the period from 2019 to 2023. It illustrates significant seismic activity in April 2019, as shown in Figure 2.

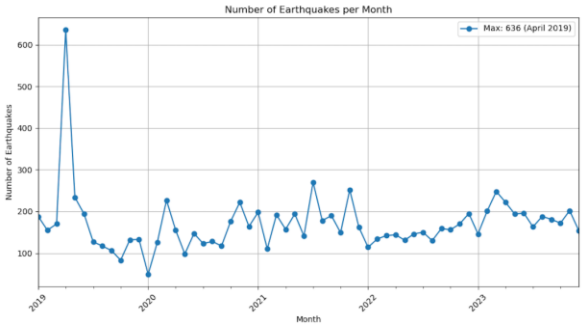

Figure 2. Earthquake distribution graph per month

The number of earthquake occurrences in the Sulawesi region exhibits a fluctuating pattern, with a noticeable increase in April 2019, followed by a range of approximately 100 to 200 events per month.

The data pre-processing provides information on the maximum, minimum, and mean values of several variables. However, we find the information from the Magnitude and Depth variables particularly interesting, as shown in Figure 3.

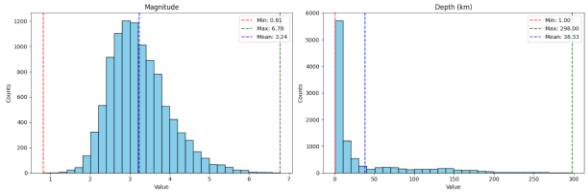

Figure 3. Graph of magnitude and depth variables

From the earthquake dataset, we can see that the average magnitude of earthquakes occurring in the Sulawesi region is 3.24 M (Magnitude), with an average depth of 38.33 km (Kilometers). This indicates that earthquakes in this region are generally of low to moderate strength but are characterized by their shallow depth. Shallow earthquakes can have a stronger impact on the Earth's surface because their epicentre is closer to the surface. This increases the risk of damage to buildings, infrastructure, and the potential for injury to residents. Therefore, despite their relatively low to moderate magnitudes, shallow earthquakes in the Sulawesi region can still significantly impact the local community and environment [15].

2.4 Tunning

To determine the optimal values for epsilon and minimum samples, a grid search method is conducted. Grid search is a systematic approach used to determine the optimal hyperparameters for machine learning algorithms, including DBSCAN clustering. In DBSCAN, the critical parameters are epsilon (eps), which defines the maximum distance between points in a cluster, and the minimum samples required to form a cluster [16].

Algorithm 1 demonstrates the grid search algorithm. This method calculates the average silhouette score across a range of epsilon parameters from 0.01 to 0.2 and a range of minimum samples from 4 to 10. Specifically, the epsilon values are tested with a step size of 0.01. This means that the algorithm evaluates epsilon values such as 0.01, 0.02, and 0.03, and continues incrementally up to 0.19, facilitating a detailed exploration within the specified range. Such granularity ensures thorough evaluation, potentially leading to the identification of the optimal epsilon value. The minimum sample values are tested as integers within this range, chosen based on empirical rules tailored to the data's dimensionality.

Algorithm 1: grid search algorithm

```
Input: Dataset X, range of ϵ values
(eps_range), range of min_samples values
(min_samples_range)
Output: Optimal ϵ (best_eps), optimal
min_samples (best_min_samples), best
evaluation score (best_score)
Initialize:
    best_eps = None
    best_min_samples = None
    best_score = -∞ (assuming a higher score is
    better, e.g., Silhouette Score)
For each ϵ in eps_range:
    For each min_samples in min_samples_range:
      Perform DBSCAN clustering with current ϵ
      and min_samples
      labels = DBSCAN(eps=ϵ,
      min_samples=min_samples).fit_predict(X)
      If the number of unique clusters in
      labels > 1:
         Calculate evaluation score for the
         current clustering (e.g., Silhouette
         Score)
         current_score = evaluate_clustering(X,
         labels)
         If current_score > best_score:
            best_score = current_score
            best_eps = ϵ
            best_min_samples = min_samples
Return best_eps, best_min_samples, best_score
Function evaluate_clustering(X, labels):
   // This function computes the evaluation
   metric, e.g., Silhouette Score
   return Silhouette Score of the clustering
```

By defining a range of values for these parameters, grid search iterates through each combination, applying DBSCAN to the dataset and evaluating the clustering quality using metrics such as the Silhouette Score. The combination of eps and minimum samples that yields the highest evaluation score is selected as the best parameter set, ensuring effective clustering.

This method provides a structured approach to parameter tuning, enhancing the robustness and accuracy of the clustering results [17]. The optimal values obtained were an epsilon of 0.06 and a minimum sample size of 9, resulting in a silhouette score of -0.06097.

2.5 Clustering

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a popular clustering algorithm used to group similar data points. It's known for its ability to find clusters of irregular shapes and identify data points that don't belong to any cluster (noise or outliers) in a dataset [18]. DBSCAN achieves this by considering how closely data points are packed together (points with many nearby neighbors) while marking points that lie alone in low-density regions (noise). It requires two parameters: epsilon (ε), which defines the maximum distance between two points to be considered neighbors, and the minimum number of points (min Pts) required to form a dense region [19].

The algorithm uses a distance metric, denoted as $d(p, q)$, which is the distance between points p and q, to determine the proximity of points. Points within the ε distance from a core point (a point with at least min Pts neighbors) are included in the same cluster. Points that do not meet this criterion are considered noise. Formula 1 illustrates the DBSCAN method.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (1)$$

Using these parameters (epsilon = 0.06 and minimum samples = 9) for DBSCAN clustering, we obtained 65 clusters.

2.6 Evaluation

The Silhouette method is a technique used to evaluate the quality of clustering by measuring how similar each data point is to its cluster compared to other clusters. Silhouette values range from -1 to 1, where a value close to 1 indicates that the data point is well-matched to its cluster and poorly matched to neighboring clusters. A value of 0 indicates that the data point is on or very close to the decision boundary between two neighboring clusters. Negative values indicate that the data point might have been assigned to the wrong cluster [20]. Formula 2 is the Silhouette calculation formula.

$$S(i) = \frac{b(i)-a(i)x^2}{max} \quad (2)$$

Where $a(i)$ is the mean distance between the data point i and all other points in the same cluster, and $b(i)$ represents the mean distance between data point i and all points in the nearest cluster (excluding i).

The Davies-Bouldin Index (DBI) is a metric used to evaluate the performance of a clustering algorithm by assessing the compactness and separation of the clusters formed. It is calculated as the average similarity ratio of each cluster with its most similar cluster. Lower DBI

values indicate better clustering performance. Specifically, for each cluster i, its similarity with another cluster j is measured by the ratio of the sum of the average distance of all points in the clusters to their respective centroids (S_i and S_j) and the distance between the centroids of the clusters (M_{ij}) [21]. Formula 3 presents the Davies Bouldin Index formula.

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{j \neq i}\left(\frac{S_i+S_j}{M_{ij}}\right) \quad (3)$$

Where k is the number of clusters, S_i is the average distance of all points in clusters i to the centroid of the cluster i, and M_{ij} is the distance between the centroid of the cluster i and j.

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, is a metric used to evaluate the quality of clustering algorithms by assessing the ratio of the sum of between-cluster dispersion and within-cluster dispersion. The higher the Calinski-Harabasz Index, the better the defined clusters are [22]. Formula 4 is Calinski-Harabasz Index formula.

$$CHIndex = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-1)} \quad (4)$$

Where $Tr(B_k)$ is the trace of the between-cluster dispersion matrix, $Tr(W_k)$ is the trace of the within-cluster dispersion matrix, n is the total number of points, and k is the number of clusters.

The silhouette score of -0.06097 suggests that the clustering may not be well-defined, as negative values typically indicate that clusters overlap or are not well-separated.

The Davies-Bouldin Index (DBI) of 2.4648, which measures the average similarity ratio of each cluster with its most similar cluster, further implies that the clusters are not very distinct; lower values are preferable for DBI.

Additionally, the Calinski-Harabasz Index, which evaluates the variance ratio between clusters and within clusters, is 138.3464. This value can be considered good. While higher values of this index generally indicate better-defined clusters, it is crucial to interpret the value within the context of the dataset and the clustering algorithm used. A higher CH Index suggests that the clusters are well-separated from each other and that the points within each cluster are compactly grouped. For DBSCAN, this index is influenced by several factors, particularly the parameters epsilon (ε) and minimum samples (min samples), which determine the density criteria for forming clusters [23].

Using various clustering evaluation metrics helps us ensure a more complete and comprehensive understanding of the quality of the generated clusters. These metrics enable us to identify issues or weaknesses in the clustering results. This approach provides a solid foundation for further improvement and optimization.

Although the evaluation results were not ideal, we proceeded with an analysis based on the identified clusters. With a total of 65 clusters, we conducted further investigation to discern spatial and temporal patterns among the earthquakes within each cluster. This analysis aims to enhance our understanding of earthquake distribution and characteristics across different parts of Sulawesi.

These insights could potentially inform disaster risk mitigation strategies and enhance infrastructure safety measures in the future. While further evaluation is necessary to refine cluster separation and clarity, the information gleaned from this clustering process remains valuable for our research efforts.

2.7 Analysis

The data collection process began with monthly data, which was then combined into annual datasets and subsequently merged into a single dataset. During data pre-processing, duplicate entries were removed based on event ID. The data was then processed using the DBSCAN algorithm. The compiled dataset identified 10,238 earthquake events in the Sulawesi region during the period from 2019 to 2023.

The DBSCAN clustering results indicate the formation of 65 clusters, The clustering also resulted in a significant amount of noise, with 3,517 data points not assigned to any cluster. Overall, these metrics suggest that the chosen parameters for DBSCAN did not yield a highly effective clustering, as evidenced by overlapping clusters and a high amount of noise.

The cluster patterns generated prompted us to attempt an analysis of the formed clusters, leading to conclusions regarding the clusters and the distribution of earthquakes in the Sulawesi region.

However, the visualization results of our analysis reveal the cluster patterns formed from the dataset, which has been processed using the DBSCAN algorithm.

3. Results and Discussions

We used the DBSCAN algorithm to identify clusters in an earthquake dataset. The Silhouette Score (-0.06097), Davies-Bouldin Index (DBI) (2.4648), and Calinski-Harabasz Index (138.3464) were used to evaluate these clusters. The Silhouette Score indicates some overlap between clusters, while the DBI suggests that the clusters are not well-separated. However, the Calinski-Harabasz Index indicates some identifiable structure in the data. Despite the complexity of seismic data, where clear separation is often challenging, this study remains informative. Our main goal is to identify seismic gaps, or areas with low seismic activity, which can still be effectively achieved despite overlapping clusters. Domain knowledge in seismology and geology further supports the interpretation of these results. Even though the clustering results may not show perfectly distinct clusters, the extracted information remains valuable for

analyzing seismic gaps and can guide further studies or disaster mitigation efforts.

We applied the DBSCAN clustering algorithm to earthquake data in the Sulawesi region for the period from 2019 to 2023. Using the resulting parameters from the grid search method with a value of epsilon = 0.06 and a minimum sample size of 9 on a dataset consisting of 10,238 records, we identified 65 clusters and detected 3,517 noise points. These results are illustrated in Figure 4.

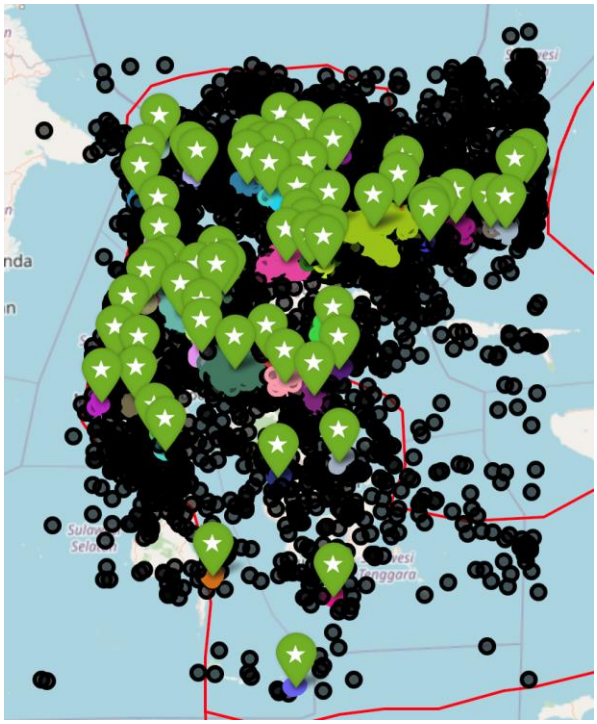

Figure 4. Result of clustering

From these results, we observe that the distribution of earthquake epicentres tends to be more numerous and spread out from the central region to the northern region.

3.1 Results

We employ distinct colour coding for each earthquake cluster in our visualization, thereby facilitating the identification of their locations and distributions. The red lines depicted in the visualization represent fault lines, providing geographic context and aiding in the comprehension of the relationship between earthquake clusters and tectonic activity in the region.

The identified clusters reveal distinct patterns of seismic activity across the Sulawesi region. High-density clusters indicate areas of increased tectonic stress, potentially linked with active faults. The variation in cluster density across the region may suggest different tectonic processes or varying levels of stress accumulation.

Interestingly, this clustering analysis also highlights several areas with lower seismic activity, which may correspond to seismic gaps. These gaps are crucial as they could be zones where stress is accumulating, potentially leading to significant earthquakes in the future. Identifying seismic gaps is essential for earthquake prediction and seismic risk mitigation. The visualization of 65 clusters without noise points is shown in Figure 5.

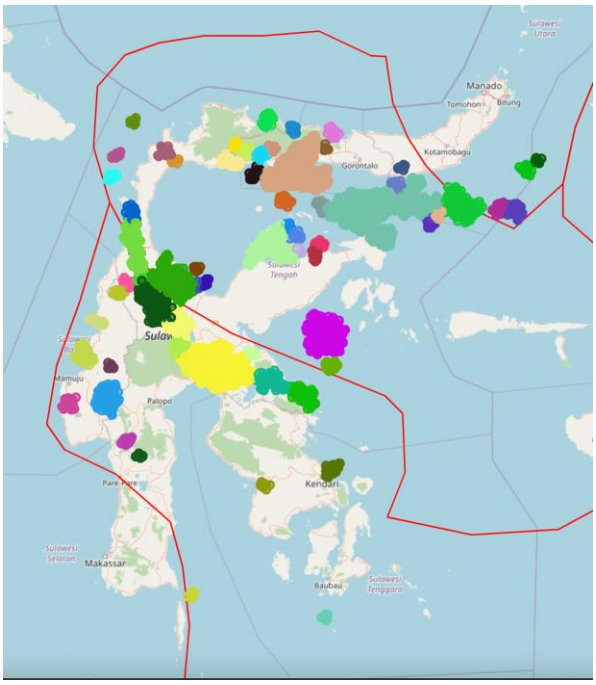

Figure 5. Cluster without noise

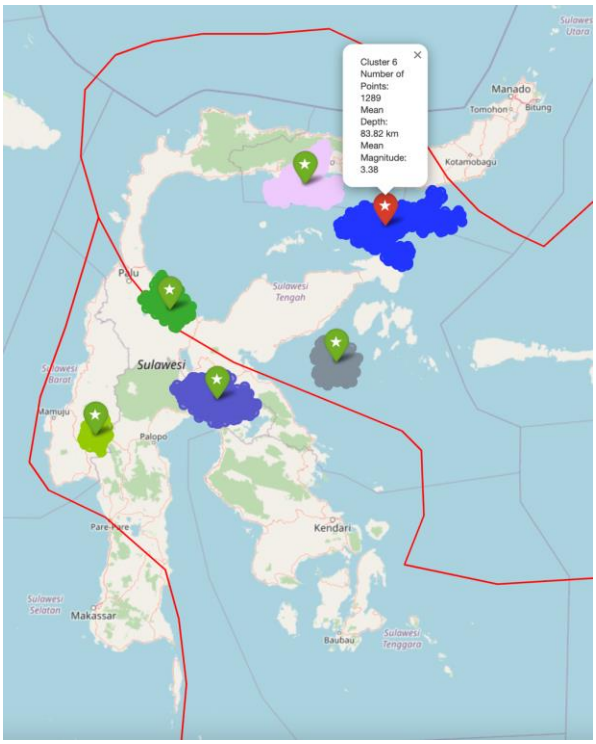

Figure 6. Top 6 Highest Cluster

From the 65 clusters formed, we can also observe areas with different density levels. We attempted to filter out six clusters with relatively high density, as shown in Figure 6. In this figure, information about the cluster

centres is displayed, with different colours indicating the cluster with the highest density. The variation in cluster density across this region may indicate different tectonic processes or varying levels of stress accumulation.

Cluster 6 is the largest cluster, with its centre located in the Molucca Sea. This cluster exhibits a relatively high average earthquake magnitude of 3.38 and an average depth of 83.82 kilometres. The presence of Cluster 6 indicates significant seismic activity in the region, particularly in the vicinity of the Manado and Gorontalo Faults.

Earthquakes with an average focal depth of 83.82 kilometres are generally categorized as deep-focus earthquakes. These earthquakes typically occur in subduction zones beneath tectonic plates, capable of affecting deep structures of the Earth while causing relatively limited surface impacts. Despite their potentially significant energy release, deep-focus earthquakes often do not cause substantial surface damage, although they can trigger tsunamis if they occur under the ocean. Buildings and infrastructure near the earthquake epicentre remain vulnerable to damage, depending on the distance and geological characteristics of the earthquake [24].

The information graph for the six largest clusters is shown in Figure 7.

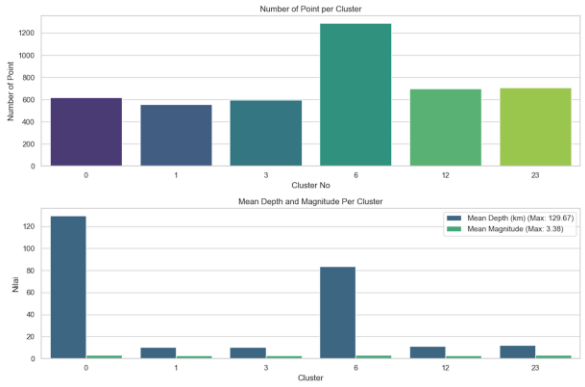

Figure 7. Top 6 Cluster Graph Mean of Magnitude and Depth

This information indicates that clusters with relatively high density exhibit significant seismic activity, with an average earthquake magnitude of 3.37 M (Magnitude), and the average depth of earthquake occurrences in those clusters is approximately 83.82 km. Table information average depth, magnitude, and the number of points for the top 6 clusters are shown in Table 2.

Based on the information from Table 2, the top six clusters exhibit significant variations in seismic activity across the Sulawesi region. Cluster No. 6 stands out with the highest number of seismic points, totalling 1289, at an average depth of approximately 83.82 km and an average magnitude of 3.38. In contrast, Cluster No. 23 shows a notable number of seismic points (709), with an average depth of 12.37 km and an average magnitude of 3.34. Cluster No. 0 exhibits the deepest average depth, reaching 129.67 km, with 618 seismic points and an average magnitude of 3.03. This cluster is located around the Southeastern Sea of Marisa, Gorontalo Province, indicating a high potential for significant tectonic activity.

Table 2. Top 6 Cluster Information Table

No	Cluster No	Number of Points	Mean Depth (km)	Mean Magnitude
1	6	1289	83.824670	3.379104
2	23	709	12.372355	3.339736
3	12	699	11.167382	2.827994
4	0	618	129.671521	3.031118
5	3	595	10.455462	2.911827
6	1	555	10.477477	2.799943

Meanwhile, other clusters are distributed around Butu Pembunian Mount (Cluster 1), Poso (Cluster 3), Faruhumfenai Nature Reserve (Cluster 12), and the western sea of Bangkurung Island (Cluster 23), each exhibiting unique characteristics in their seismic activity. This information provides valuable insights for further understanding the potential earthquake risks and seismic risk management in the surrounding areas.

3.2 Discussion

An analysis of earthquake depths in Clusters 6 and 0 reveals crucial insights into their potential impacts. Cluster 0, exhibiting the deepest average depth of 129.67 km, suggests the presence of deep-focus earthquakes. Deep-focus earthquakes generally cause less surface damage compared to shallow ones due to the dissipation of seismic energy over a greater volume of rock. However, they can be felt over larger areas and may indicate significant tectonic activities at depth, which could be precursors to more substantial seismic events [25].

Therefore, understanding these depth-related characteristics is vital for assessing the potential earthquake risks in the Sulawesi region and developing effective seismic risk management strategies. These insights underscore the importance of continuous monitoring and analysis to anticipate and mitigate the impacts of seismic activities.

According to several studies, seismic gaps are segments of tectonic faults that have not experienced significant earthquakes for a prolonged period, despite surrounding areas being seismically active. These regions are considered to have a high potential for large future earthquakes due to the continuous accumulation of tectonic stress [26]. The identification of seismic gaps is crucial for earthquake prediction and seismic risk assessment [27].

Identifying high-density seismic clusters can provide local authorities and policymakers with crucial information about the most at-risk areas. By focusing preparedness efforts on these regions, it is possible to enhance community resilience and mitigate the potential impacts of future earthquakes. Additionally, recognizing seismic gaps can guide further geological

investigations to better understand and monitor these critical areas [28].

A significant concentration of earthquakes is observed in the southeastern region of Kotamubago, North Sulawesi Province. Notably, substantial earthquake activity has been identified in the vicinity of the Molucca Sea.

However, several small clusters have formed southeastward and near the North Sulawesi subduction zone. The North Sulawesi subduction zone is an area where the Pacific Plate subducts beneath the Eurasian Plate, creating complex tectonic conditions that often lead to earthquakes and volcanic activity in the region [29].

The interaction between these plates influences the geological and seismic characteristics, enhancing our understanding of earthquake hazards in this area. Figure 8 shows the distribution of clusters and noise in North Sulawesi.

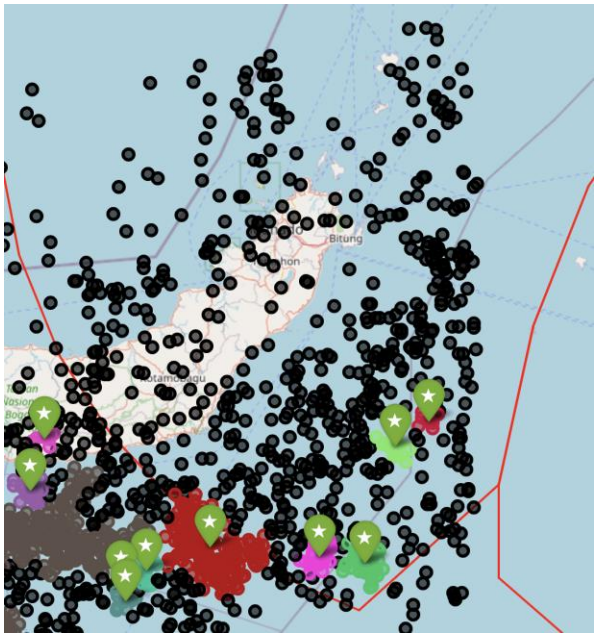

Figure 8. Cluster in North Sulawesi

Although there are no clusters formed in the North Sulawesi region, this area exhibits significant seismic activity. The clustering results indicate that earthquakes occur scattered rather than centred at a single point. Therefore, we believe that areas not yet identified as epicentres have a high potential to become future epicentres.

In the Gorontalo province area, small clusters are observable with two major clusters identified, surrounding the city of Gorontalo, which is bordered by small clusters, potentially rendering Gorontalo as a seismic gap area based on the clustering pattern as shown in Figure 9.

The surrounding existing clusters suggest that the city of Gorontalo itself has the potential to become a new cluster in the future. This highlights the possibility that Gorontalo is currently a seismic gap area but also a potential future hotspot for seismic activity.

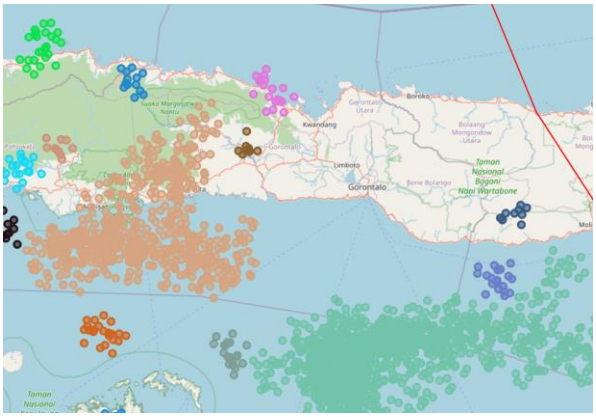

Figure 9. Cluster in Gorontalo

Central Sulawesi Province encompasses a vast area. After the earthquake in 2018 that struck Palu City, we pinpointed relatively dense seismic activity around the vicinity of Palu, with eight identified clusters formed.

This high density of clusters can be attributed to the active Palu-Koro fault, a major fault line that runs through Central Sulawesi. The Palu-Koro fault is a significant tectonic boundary where the Australian and Eurasian plates interact, leading to frequent and intense seismic activity [30]. The clustering pattern in Central Sulawesi Province helps identify potential seismic gap areas. Notably, Poso, Ampana, and Morowali appear to be located between three clusters exhibiting significant seismic activity. This suggests these regions could be seismic gaps.

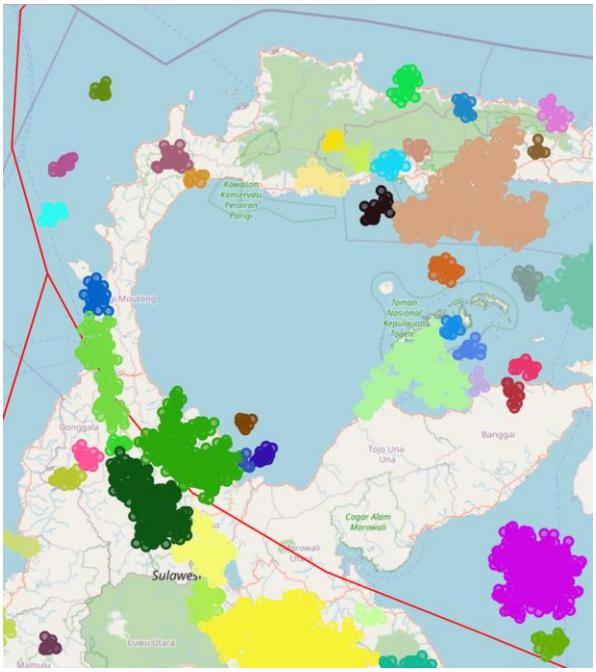

Figure 10. Central Sulawesi cluster

Additionally, the Toli-Toli and Buol areas may also become future seismic hotspots due to their proximity to smaller clusters. Figure 10 illustrates the distribution

of clusters in Central Sulawesi. Given the paramount importance of mitigating earthquake disasters posed by Palu-Koro Fault activities, effective mitigation measures are essential. These measures can include strengthening building structures, enhancing early warning systems, and educating the public on pre-earthquake, during-earthquake, and post-earthquake actions. Recent research suggests that risk mapping and probabilistic analysis of past seismic activity can be instrumental in designing more effective mitigation strategies. Therefore, a thorough understanding of the Palu-Koro Fault dynamics and the implementation of appropriate mitigation measures are crucial steps towards reducing the impact of future earthquake disasters [31]. In West Sulawesi province, there is an indication of seismic gap potential in the Mamuju area. Figure 11 displays the cluster formation in the West Sulawesi area.

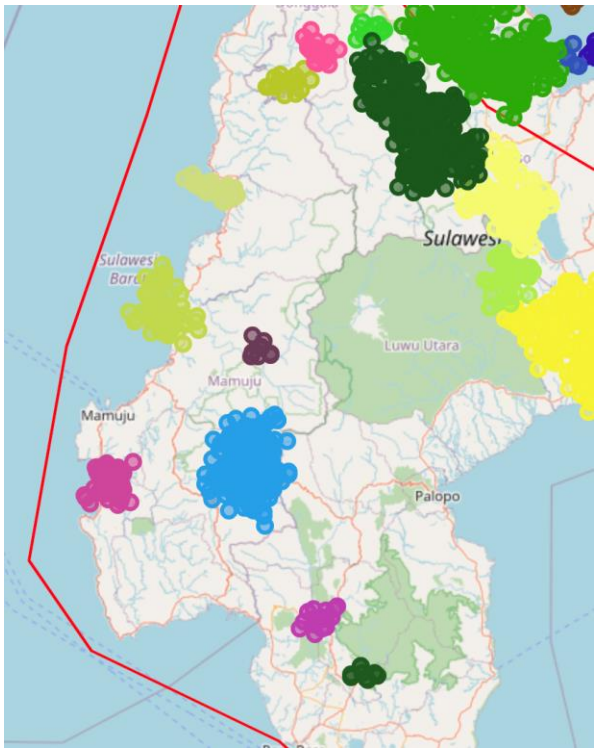

Figure 11. Cluster in West Sulawesi

The distribution of earthquakes in the western province forms small clusters, with epicentres and clusters tending to spread towards the Central Sulawesi border. We believe this is due to seismic activity along the Palu-Koro fault in Central Sulawesi [32]. However, the small clusters formed in West Sulawesi Province strongly indicate seismic gap areas within this province. Many small areas are surrounded by these earthquake clusters. Seismic activities in Southeast Sulawesi province appear significant in the clustering pattern around Kendari and Kolaka cities. In this region, earthquake epicentres during the period 2019-2023 show dispersion but do not form many clusters. The dispersion of earthquake epicentres in the Southeast Sulawesi region is evident in Figure 12.

Figure 12. Cluster in Southeast Sulawesi

Earthquake epicentres are quite spread out towards the north in Southeast Sulawesi, resulting in a few clusters forming in this region. However, this area includes small islands where, according to clustering data from 2019-2023, no earthquake clusters have formed. The identified clusters and epicentres in this region tend to be located towards the north, near Central Sulawesi Province.

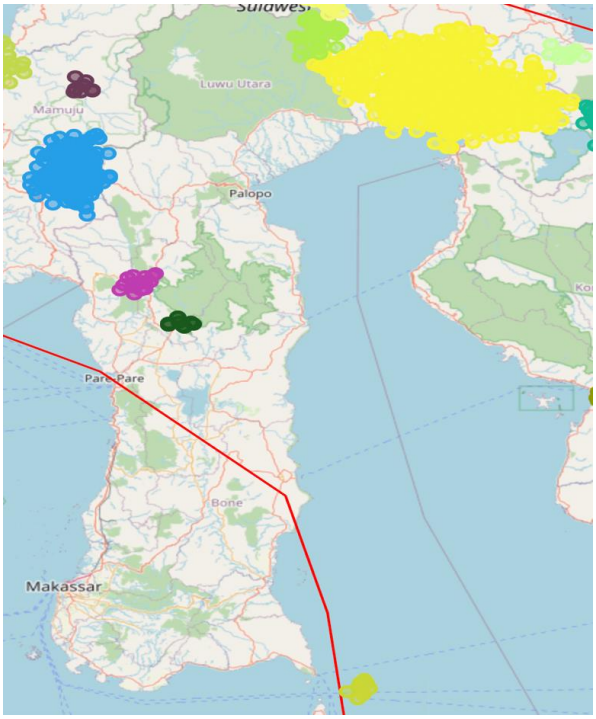

Figure 13. Cluster in South Sulawesi

Five clusters have been identified in South Sulawesi province, with the largest cluster located in the

Morowali area near the border with Central Sulawesi province. The clustering pattern and dispersion of earthquake epicentres in the South Sulawesi region are shown in Figure 13.

Around the city of Makassar, earthquake clusters may not have formed, but the epicentres in South Sulawesi Province tend to be more numerous in the western region and towards the north, near the Central Sulawesi border. However, upon closer examination, earthquake epicentres are also spread throughout this southern province, although clusters may not have formed yet.

Analyzing the clustering patterns annually from the Sulawesi earthquake dataset spanning 2019 to 2023 provides profound insights into the evolving seismic dynamics over time. Tracking how earthquakes organize into clusters each year enables the identification of significant spatial distribution trends. For instance, the formation of clusters can indicate areas in Sulawesi prone to intensive seismic activity, whereas regions lacking clusters may highlight potentially seismically quiet or undetected areas.

The identification of annual clustering patterns indicates an increase in earthquake epicentres or distribution between central and northern Sulawesi. Each year, Central Sulawesi consistently shows a significant level of clustering. Figure 14 illustrates the variations in clustering patterns across Sulawesi over the years.

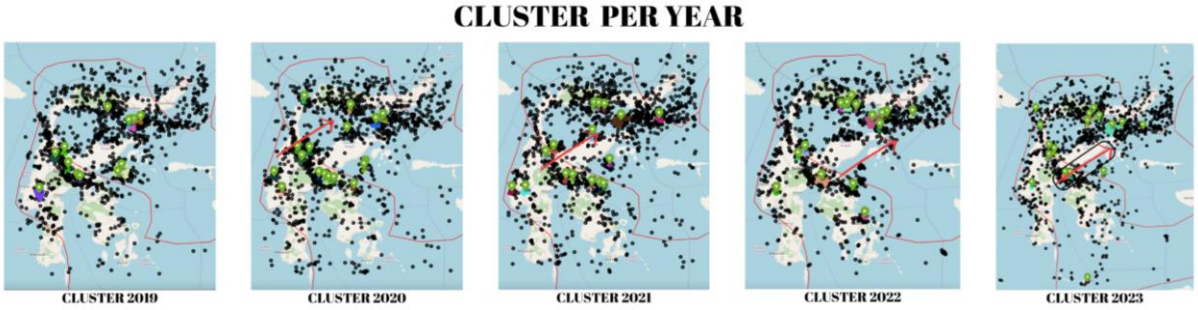

Figure 14. Cluster pattern per year

Based on the table summarizing earthquake data from 2019 to 2023, the number of earthquakes fluctuated each year, reaching peaks of 2275 in 2019 and 2267 in 2023, with a low of 1732 in 2020. The total number of clusters varied, peaking at 17 in both 2020 and 2023, and reaching a low of 12 in 2019. Noise (non-clustered events) showed an increasing trend, with the highest level recorded in 2023 (1454) and the lowest in 2019 (1168). The average earthquake magnitude slightly decreased from 3.3 in 2019 and 2020 to 3.1 in 2023, while the average earthquake depth steadily increased from 28.06 km in 2019 to 44.81 km in 2023. The silhouette score, which measures clustering quality, consistently remained negative, indicating poor clustering quality, with the lowest score recorded in 2020 (-0.30644) and the highest in 2019 (-0.09022). Overall, despite the increase in the number of earthquakes and noise, the average magnitude slightly decreased, and earthquake depth increased. The poor clustering quality throughout the period is likely due to the wide distribution of earthquakes and significant variations in depth, which reduce the uniformity and cohesion of the clusters formed [33]. Table 3 shows the information on clusters per year.

Table 3. Information of cluster per year

Year	Total of Earthquake	Total of Cluster	Total of Noise	Mean Magnitude	Mean Depth (km)	Silhouette Score
2019	2275	12	1168	3.3	28.06	-0.09022
2020	1732	17	1210	3.3	36.41	-0.30644
2021	2193	16	1336	3.2	40.00	-0.15401
2022	1771	15	1141	3.2	43.01	-0.25860
2023	2267	17	1454	3.1	44.81	-0.29070

The analysis of the data reveals a notable increase in the average earthquake depth from 2019 to 2023. While the average magnitude remained stable between 3.1 and 3.3, the depth steadily rose from 28.06 km in 2019 to 44.81 km by 2023. This upward trend suggests a shift in seismic behaviour, with earthquakes occurring at deeper levels. Understanding these changes is crucial for effective disaster risk mitigation and safeguarding infrastructure in the future.

4. Conclusions

The analysis of Sulawesi earthquake data from 2019 to 2023 using the DBSCAN method (epsilon: 0.06, minimum sample size: 9) identified 65 earthquake clusters, mostly shallow. Significant areas of potential seismic gaps include Manado, Gorontalo, and Buol. Evaluation metrics show room for improvement: the Silhouette Score is -0.06097, the Davies-Bouldin Index is 2.4648, and the Calinski-Harabasz Index is 138.3464, indicating moderate cluster quality. These results

suggest the need for further refinement to achieve better cluster separation. Future research should focus on fine-tuning the DBSCAN parameters (epsilon and minimum sample size) to improve cluster identification. Analyzing earthquake characteristics based on depth, especially shallow earthquakes (20 km to 50 km depth), is crucial due to their potential for significant damage. Additionally, studying the annual increase in earthquake clusters and developing disaster preparedness strategies will be important. Fine-tuning clustering parameters based on local geological characteristics and earthquake distribution patterns will be essential for obtaining more precise and reliable results in future studies.

Acknowledgements

We would like to thank Dr. Hadi Santoso, S. Kom., M. Kom., and Mr. Lukman Hakim, S.T., M. Kom., for their valuable contributions to this research. Their insights and suggestions were instrumental in shaping this study. We also thank BMKG and the Faculty of Computer Science at Universitas Mercu Buana Jakarta for their support.

References

[1] S. J. Hutchings and W. D. Mooney, "The Seismicity of Indonesia and Tectonic Implications," *Geochemistry, Geophysics, Geosystems*, vol. 22, no. 9, Sep. 2021, doi: 10.1029/2021GC009812.

[2] A. Bobbette, R. Gamble, C. T. Lee, and C. Wilson, "Decolonizing Geology: A Discussion," *GeoHumanities*, vol. 7, no. 2, pp. 647–655, 2021, doi: 10.1080/2373566X.2021.1896373.

[3] Koesnama, "Pensesaran Mendatar Dan Zona Tunjaman Aktif Di Sulawesi: Hubungannya Dengan Kegempaan," *Pensesaran Mendatar Dan Zona Tunjaman Aktif Di Sulawesi: Hubungannya Dengan Kegempaan*, vol. 15, 2014.

[4] A. Amalia, U. Harmoko, and G. Yuliyanto, "Clustering of Seismicity in the Indonesian Region for the 2018-2020 Period using the DBSCAN Algorithm," 2021. [Online]. Available: https://ejournal2.undip.ac.id/index.php/jpa/index

[5] A. Putri, W. Hadi, H. Pratiwi, and I. Slamet, "Pengelompokan Data Gempa Bumi di Indonesia dengan Algoritma K-Means dan DBSCAN," 2023.

[6] M. Bariklana and A. Fauzan, "Implementation Of The Dbscan Method For Cluster Mapping Of Earthquake Spread Location," *Barekeng: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 2, pp. 0867–0878, Jun. 2023, doi: 10.30598/barekengvol17iss2pp0867-0878.

[7] B. Maruli Siahaan and A. Roma Rio, "Agglomerative Clustering of 2022 Earthquakes in North Sulawesi, Indonesia," *Buana Information Technology and Computer Sciences (BIT and CS*, vol. 4, no. 2, p. 77, 2023, [Online]. Available: https://repogempa.bmkg.go.id/

[8] S. Harini, H. Fahmi, A. D. Mulyanto, and M. Khudzaifah, "The earthquake events and impacts mapping in Bali and Nusa Tenggara using a clustering method," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics Publishing, Apr. 2020. doi: 10.1088/1755-1315/456/1/012087.

[9] P. Bountzis, E. Papadimitriou, and G. Tsaklidis, "Identification and Temporal Characteristics of Earthquake Clusters in Selected Areas in Greece," *Applied Sciences (Switzerland)*, vol. 12, no. 4, Feb. 2022, doi: 10.3390/app12041908.

[10] Y. Rong, D. D. Jackson, and Y. Y. Kagan, "Seismic gaps and earthquakes," *J Geophys Res Solid Earth*, vol. 108, no. B10, Oct. 2003, doi: 10.1029/2002jb002334.

[11] A. Wahyu and Rushendra, "Klasterisasi Dampak Bencana Gempa Bumi Menggunakan Algoritma K-Means di Pulau Jawa," *JEPIN*, vol. 8, no. 1, 2022.

[12] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, Sep. 2017, doi: 10.3923/jeasci.2017.4102.4107.

[13] K. M. Arsyad, A. Yunita, H. Mas'uudah Krismartopo, A. Syahputri Dimar, K. Dewi, and I. Madrinovella, "Revealing Insights Through Exploratory Data Analysis on Earthquake Dataset."

[14] L.-P. Chen, "Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python," *Technometrics*, vol. 63, no. 2, pp. 272–273, Apr. 2021, doi: 10.1080/00401706.2021.1904738.

[15] H. F. Yang and S. Yao, "Shallow destructive earthquakes," *Earthquake Science*, vol. 34, no. 1. Earthquake Science, pp. 15–23, 2021. doi: 10.29382/EQS-2020-0072.

[16] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," 2011. [Online]. Available: http://scikit-learn.sourceforge.net.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996. [Online]. Available: www.aaai.org

[18] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," *Data Knowl Eng*, vol. 60, no. 1, pp. 208–221, Jan. 2007, doi: 10.1016/j.datak.2006.01.013.

[19] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems*, vol. 42, no. 3, Jul. 2017, doi: 10.1145/3068335.

[20] P. J. Rousseeuw, "Silhouet tes: a graphic al aid to the interpre tation and validati on of cluster analysis," 1987.

[21] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans Pattern Anal Mach Intell*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[22] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, Jan. 1974, doi: 10.1080/03610927408827101.

[23] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-Based Clustering Based on Hierarchical Density Estimates."

[24] A. Kusmiran, Minarti, M. F. I. Massinai, A. Zarkasi, A. A. Maharani, and R. Desiani, "Klasifikasi Kedalaman Kejadian Gempa Menggunakan Algoritma K-Means Clustering: Studi Kasus Kejadian Gempa Di Sulawesi," *JFT: Jurnal Fisika dan Terapannya*, vol. 9, no. 2, pp. 79–88, Dec. 2022, doi: 10.24252/jft.v9i2.29198.

[25] Z. Zhan, "Mechanisms and Implications of Deep Earthquakes," 2019, doi: 10.1146/annurev-earth-053018.

[26] R. Guo, Y. Zheng, and J. Xu, "Stress modulation of the seismic gap between the 2008 Ms 8.0 Wenchuan earthquake and the 2013 Ms 7.0 Lushan earthquake and implications for seismic hazard," *Geophys J Int*, vol. 221, no. 3, pp. 2113–2125, 2020, doi: 10.1093/GJI/GGAA143.

[27] W. Thatcher, "Earthquake recurrence and risk assessment in circum-Pacific seismic gaps," *Nature*, vol. 341, no. 6241, pp. 432–434, 1989, doi: 10.1038/341432a0.

[28] M. Ibrahim and B. Al-Bander, "An integrated approach for understanding global earthquake patterns and enhancing seismic risk assessment," *International Journal of Information Technology (Singapore)*, vol. 16, no. 4, pp. 2001–2014, Apr. 2024, doi: 10.1007/s41870-024-01778-1.

[29] T. Zheng, Q. Qiu, J. Lin, and X. Yang, "Raised potential earthquake and tsunami hazards at the North Sulawesi subduction zone after a flurry of major seismicity," *Mar Pet Geol*, vol. 148, Feb. 2023, doi: 10.1016/j.marpetgeo.2022.106024.

[30] D. H. Natawidjaja *et al.*, "The 2018 Mw7.5 Palu 'supershear' earthquake ruptures geological fault's multisegment separated by large bends: Results from integrating field measurements, LiDAR, swath bathymetry and seismic-

reflection data," *Geophys J Int*, vol. 224, no. 2, pp. 985–1002, Feb. 2021, doi: 10.1093/gji/ggaa498.

[31] S. Pasari, A. V. H. Simanjuntak, Neha, and Y. Sharma, "Nowcasting earthquakes in Sulawesi Island, Indonesia," *Geoscience Letters*, vol. 8, no. 1. Springer Science and Business Media Deutschland GmbH, Dec. 01, 2021. doi: 10.1186/s40562-021-00197-5.

[32] T. Kiyota, H. Furuichi, R. F. Hidayat, N. Tada, and H. Nawir, "Overview of long-distance flow-slide caused by the 2018 Sulawesi earthquake, Indonesia," *Soils and Foundations*, vol. 60, no. 3, pp. 722–735, Jun. 2020, doi: 10.1016/j.sandf.2020.03.015.

[33] M. D. Petersen *et al.*, "Documentation for the 2014 Update of the United States National Seismic Hazard Maps", [Online]. Available: http://dx.doi.org/10.3

