Journal of Current Studies in SDGs
Homepage : https://journal.i-ros.org/index.php/JOCSIS
Email : jocsis@i-ros.org

e-ISSN 3089-7564
Vol.2, No.2, June 2026

Data-Driven Seismic Hazard Zonation of Indonesia to Support SDG 11 Using DBSCAN and K-Means Clustering

Atik Wintarti, Fadhilah Qalbi Annisa, Harmon Prayogi*, Yuliani Puji Astuti, and Ibnu Febry Kurniawan

Universitas Negeri Surabaya, Surabaya, Indonesia

DOI : https://doi.org/10.63230/jocsis.2.2.224

Sections Info

Article history:
Submitted: May 27, 2026
Final Revised: June 4, 2026
Accepted: June 5, 2026
First Available Online: June 26, 2026
Publication Date: June 27, 2026

Keywords:
DBSCAN;
K-Means;
Machine Learning;
Seismic Hazard Zonation.

ABSTRACT

Objective: *To examine ten years of earthquake data recorded across Indonesia drawing on 5,364 events with magnitudes above M 5.0 between 2016 and 2025.* ***Method:*** *DBSCAN algorithm was run after the optimal neighborhood radius was determined objectively from a k-distance plot. An elbow at about 65 km was identified and the value yielded 16 spatially distinct clusters alongside 460 noise events. K-means algorithm identified four seismic regimes.* ***Results:*** *Of the four regimes, one cluster (Cluster 1) concentrated every major earthquake in the catalog (64 events with M >= 7.0), even though it accounted for fewer than one event in ten. The three remaining clusters captured background seismicity at near-identical mean magnitudes of approximately from 5.33 to 5.35. At the conventional zonal level, Maluku-Sulawesi generated the most events about 40.8% from total events, while Sumatra registered the highest seismic energy output. A Gutenberg-Richter b-value of 0.98 was estimated for the full catalog.* ***Novelty:*** *Introducing earthquake zonation methods based on machine learning for earthquake catalog of Indonesia. These findings support multiple Sustainable Development Goals including the identification of underestimated high-energy rupture corridors informs evidence-based urban risk reduction (SDG 11), strengthens the scientific foundation for earthquake disaster preparedness (SDG 13), introduces an innovative and reproducible machine learning methodology applicable to infrastructure (SDG 9), and contributes a freely transferable workflow that adopt data-driven zonation methods (SDG 17).*

INTRODUCTION

The 2004 Sumatra-Andaman rupture reached Mw 9.1-9.3 and set off a tsunami that killed more than 200,000 people across fourteen countries (Sarkawi et al., 2024). The 2009 Padang earthquake (Mw 7.6) caused widespread structural collapse across West Sumatra (Kadir et al., 2024; Natawidjaja & Triyoso, 2007). In 2018, the Palu earthquake and liquefaction disaster demonstrated that even moderate-length fault segments can trigger catastrophic cascading failures when soil conditions and coastal geometry conspire against a population (Cilia et al., 2022; Heidarzadeh & Mulia, 2023). These repeated earthquake phenomena share a common thread including communities and infrastructure. They were exposed to levels of ground shaking that hazard assessments had either underestimated or communicated inadequately to those responsible for land-use planning and building codes (Iuchi et al., 2023; Stein & Wysession, 2003). Conventional seismic hazard zonation addresses this problem by dividing a study region into geographic provinces that estimate earthquake recurrence within each province and translating those estimates into design ground motions. The approach is well-established and has produced the national seismic hazard map of Indonesia (PusGen, 2017). However, its limitation is that zone boundaries are drawn by experts exercising professional judgment, and there is no guarantee that those boundaries correctly capture the underlying structure of seismicity.

Indonesia's destructive earthquakes have motivated decades of seismic hazard research at national and regional scales. The most authoritative national-level product is the 2017 Seismic Hazard Map of Indonesia, developed by the National Center for Earthquake Studies (PusGen, 2017) and formally documented by Irsyam et al. (2020) who incorporated updated source models for active shallow crustal faults, subduction interfaces, and background seismicity, together with new ground motion prediction equations and site amplification corrections. Regional studies have uncovered seismogenic structures and hazard gaps not captured in the 2017 national map, including fault locking along the Baribis fault south of Jakarta, revised ground motion estimates for the greater Jakarta area, hazard implications of the 2022 Pasaman earthquake sequence in West Sumatra, and the identification of the previously unmapped Cugenang Fault as the causative structure of the destructive 2022 Cianjur earthquake (Widiyantoro et al., 2022; Damanik et al., 2023; Wulandari et al., 2023; Zulfakriza et al., 2024; Jihad et al., 2021; Marliyani et al., 2016; Widiyantoro & van der Hilst, 1997).

The b-value of the Gutenberg-Richter relation has been used widely as a proxy for differential stress and seismogenic potential across Indonesian tectonic domains (Schorlemmer et al., 2005; Sari et al., 2023; R & Madrinovella, 2024; Wajedy et al., 2026). These studies collectively confirm that b-value spatial variation is a meaningful seismotectonic indicator across Indonesian tectonic zones, providing independent validation for the zonal differences in Gutenberg-Richter parameters reported in the present study.

The 2018 Central Sulawesi earthquake sequence has been particularly influential in reshaping understanding of cascading hazard processes. Supendi et al. (2020) relocated aftershocks from both the 2018 Lombok and Palu sequences that clarify the geometry of the Palu-Koro fault rupture and its relationship to background seismicity in eastern Indonesia. Field observations from the Palu event documented ground shaking intensities reaching Modified Mercalli Intensity alongside extensive liquefaction and tsunami runup, a dual-source tsunami model combining fault slip and submarine landslide contributions revealed that conventional single-source approaches underestimate near-field wave heights, post-disaster analysis showed that hazard map revisions did not fully translate into risk-informed land-use decisions, and the broader context of Indonesian subduction seismicity is framed within the global record of megathrust ruptures (Cilia et al., 2022; Heidarzadeh & Mulia, 2023; Iuchi et al., 2023; Bilek & Lay, 2018).

In the domain of earthquake risk assessment, Jena et al. (2020) applied machine learning approaches including clustering analysis to identify earthquake-prone areas at Palu, Indonesia, integrating earthquake probability, susceptibility to seismic amplification, and structural vulnerability into composite risk maps. Their work represents an early application of data-driven spatial analysis to Indonesian hazard assessment and demonstrates the feasibility of combining zonation based on Machine Learning (ML) with traditional hazard products. More recently, Kadir et al. (2024) reviewed the evolution of Indonesian seismic load codes from SKBI 1987 through the current SNI standards documenting how successive revisions have progressively increased design spectral accelerations in response to improved hazard understanding following major earthquakes.

Machine Learning offers a data-driven complement to expert-defined zonation. Unsupervised clustering algorithms impose no predetermined boundaries (Seydoux,

2020; Vijay & Nanda, 2023). Instead, they search for groups within the data itself guided purely by the statistical properties of the observations. DBSCAN discovers clusters by measuring the local density of points which regions where many events fall close together in space are assigned to clusters while isolated events are labelled noise. K-means partitions events by minimizing the total within-group variance across a specified number of groups which allows magnitude to be incorporated alongside coordinates so that groups reflect seismic energy character as well as location (Sabermahani & Frederiksen, 2024; Katta et al., 2024).

The study domain extends from 11 degrees South to 6 degrees North latitude and from 95 to 141 degrees East longitude covering the full geographic extent of the Indonesian archipelago. Within this area, three major plate interactions drive seismicity. In the west and center, the Indo-Australian Plate descends beneath the Eurasian Plate along the Sunda Trench, a subduction zone stretching more than 5,600 km from the Andaman Sea through Java to Timor (Curray, 2005; McNeill et al., 2014). Convergence rates along this boundary approach 70 mm per year, and the resulting megathrust interface has generated some of the largest earthquakes in the instrumental record (Tregoning et al., 1994; McCaffrey, 2009).

Eastern Indonesia operates under a more fragmented tectonic regime. Sulawesi locates at the intersection of at least three microplates, and its seismicity reflects the resulting complexity of stress orientations and fault geometries (Socquet et al., 2006; Hutchings & Mooney, 2021). The Maluku Sea is bounded by two subduction zones dipping away from each other, a configuration known as a double subduction system that concentrates seismogenic potential in a relatively small geographic area (McCaffrey, 1982; Zhang et al., 2017). Western New Guinea accommodates oblique collision between the Australian craton and island-arc terranes of the Pacific realm, generating both thrust and strike-slip seismicity across the fold-and-thrust belt of the central highlands (Abers & McCaffrey, 1988).

This study applies both algorithms to a ten-year earthquake catalog, with eps parameters selected objectively from a k-distance plot rather than set arbitrarily, and with the K-means partition count optimized through Elbow and Silhouette measurements. The goals include to characterize background seismicity across the five conventional tectonic zones, to apply DBSCAN to identify spatially coherent seismogenic segments and isolate anomalous events, to use K-means to separate seismic regimes by their combined spatial and energy characteristics, and to compare the two frameworks and discuss what each contributes to earthquake hazard understanding.

RESEARCH METHOD

Earthquake catalog

Earthquake records were obtained from the IRIS WILBER 3 seismic database from the period of January 2016 to December 2025. After filtering to retain only events with M > 5.0 occurring within Indonesian territory. The study area is divided into five zones delimited by longitude including Sumatra (west of 108 degrees E), Java (108 to 116 degrees E), Bali-Nusa Tenggara (116 to 122 degrees E), Maluku-Sulawesi (122 to 130 degrees E), and Western New Guinea (east of 130 degrees E). The working catalog contains 5,364 records. Each record carries four attributes: a sequential index, hypocenter latitude, hypocenter longitude, and moment magnitude. No focal depth information was available in the dataset, so all spatial analyses are conducted in two geographic dimensions. The M 5.0 completeness threshold was chosen because network

detectability across the broad Indonesian monitoring domain is considered near-uniform above this level.

Conventional zone analysis

Each event was assigned to one of the five longitude-defined zones, and within each zone the following statistics were computed such as total event count and its share of the catalog, mean magnitude with standard deviation, maximum recorded magnitude, and the count of major events defined as those with M≥7.0. A composite hazard score was derived from normalized values of frequency, maximum magnitude, and major-event count, weighted 0.35, 0.40, and 0.25, respectively. The Gutenberg-Richter frequency-magnitude relation (log N equals a minus b times M) was fitted to the full catalog and to each zone separately by maximum likelihood estimation, and fitness was measured by the coefficient of determination R-squared (MacQueen, 1967; Gutenberg & Richter, 1944).

DBSCAN clustering

DBSCAN was applied to the geographic coordinates using the Haversine distance metric which correctly accounts for the curvature of the Earth and is therefore more appropriate than Euclidean distance at the spatial scales involved in this study. The algorithm requires two parameters such as eps that is the maximum distance within which two points are considered neighbors, and minPts that is the minimum number of neighbors a point must have to qualify as a core point. Setting minPts to 15 reflects a judgment that seismogenic clusters should contain at least fifteen events to be considered geophysically meaningful.

The value of eps parameter was not assumed but determined from the data through a k-distance plot. For each event, the distance to its fifth nearest neighbor (k = 5, matching minPts) was computed and sorted from largest to smallest. When this curve is plotted against rank, well-clustered data produce a steep elbow including below the elbow, distances are small and points belong to dense clusters. The elbow was located using a maximum-curvature criterion that finds the point farthest from the diagonal of the normalized plot. This procedure yielded eps = 0.59 degrees, this corresponding to approximately 65 km at Indonesian latitudes. All DBSCAN computations were performed using the BallTree algorithm within scikit-learn Python module (Pedregosa et al., 2011).

K-means clustering

K-means was applied to three features such as latitude, longitude, and magnitude of events in the earthquake catalog. Because these variables differ in scale and physical units, all three were standardized using z-score normalization before clustering. Without this step, the geographic coordinates would numerically dominate the magnitude variable which spans roughly four units, and K-means would effectively ignore magnitude when forming groups.

The number of clusters K was selected by evaluating two independent diagnostics across the range from K = 2 to K = 11. The Elbow criterion plots within-cluster sum of squared errors against K. The point where the rate of decrease slows most sharply suggests the appropriate partition count. The Silhouette Score measures how similar each event is to its own cluster compared to the nearest alternative cluster with values closer to positive one indicating better-separated groups (Rousseeuw, 1987). Ten

independent random initializations were used to reduce sensitivity to the starting configuration, and a fixed random seed was applied to guarantee reproducibility.

RESULTS AND DISCUSSION

Catalog overview

The mean magnitude is 5.44 and the standard deviation is 0.41 reflecting the dominance of moderate events. As shown in Figure 1, seismicity is concentrated along the Sunda Trench offshore Sumatra and Java, the Banda Arc through Bali and the Lesser Sunda Islands, and the multiply faulted crust of Sulawesi and Maluku. Interior Borneo is almost entirely absent from the record.

The magnitude frequency histogram shown in Figure 2a follows a classic exponential pattern which the moderate class (M 5.0-5.4) accounts for 73.7% of all events, while the four events that reached M≥8.0 represent less than 0.1% of the catalog. The Gutenberg-Richter fit in Figure 2b is excellent with R-squared = 0.996 returning a b-value of 0.98 and an a-value of 8.66, this values consistent with the subduction-dominated character of Indonesian seismicity (Hutchings & Mooney, 2021; Irsyam et al., 2020).

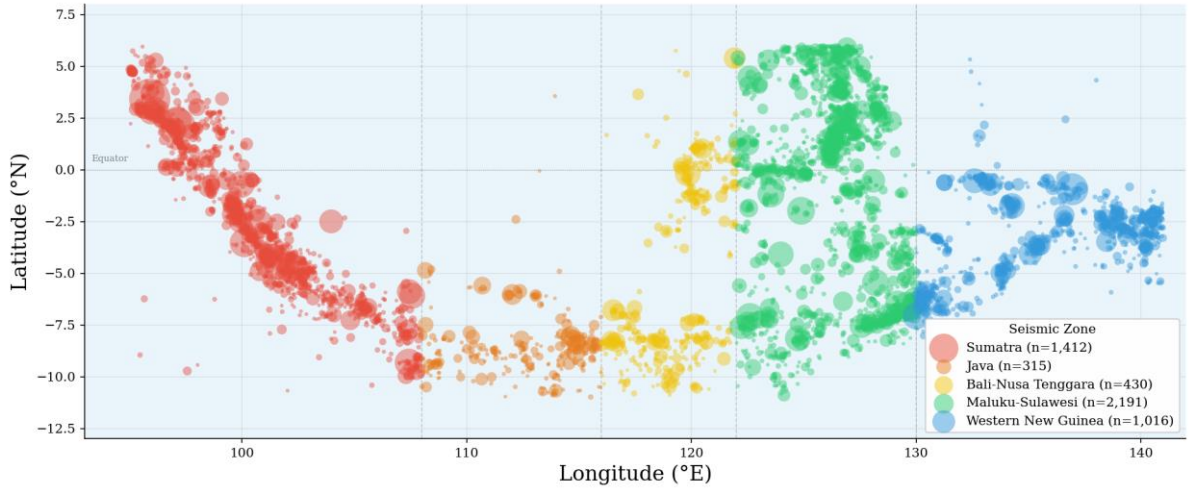

Figure 1. Spatial distribution of M≥5.0 earthquakes in Indonesia. Symbol size is proportional to magnitude; colors identify the five conventional seismic zones. The three largest events are annotated

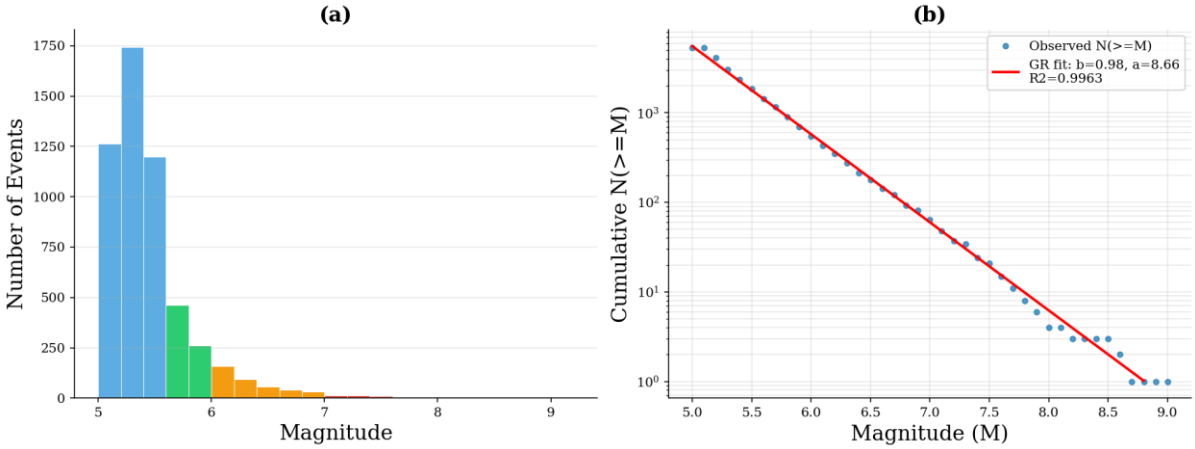

Figure 2. (a) Magnitude distribution histogram with color-coded magnitude classes. (b) Gutenberg-Richter cumulative frequency-magnitude plot with best-fit line b = 0.98 and R-squared = 0.996

Zone-level statistics

Table 1 presents the seismicity statistics for the five conventional zones. From all 2,191 events, the Maluku-Sulawesi zone contributed more earthquakes than any other zone, a result of the concentrated multi-plate interactions that characterize eastern Indonesia. Sumatra was second in frequency with 1,412 events, but it produced both the highest single magnitude (M 9.0) and the greatest composite hazard score. Western New Guinea recorded an M 8.1 event, the second largest in the catalog, consistent with the compressional stress field generated by Australian-Pacific plate convergence. Mean magnitudes across all five zones are remarkably uniform ranging only from 5.43 to 5.46.

Table 1. Seismicity statistics by conventional seismic zone

Zone	Events	Events Percentage	Mean M	Max M	$M \geq 7.0$ Events	Hazard Level
Sumatra	1,412	26.3%	5.45	9.0	20	High
Maluku-Sulawesi	2,191	40.8%	5.44	7.7	27	High
Western New Guinea	1,016	18.9%	5.43	8.1	12	Moderate-High
Bali-Nusa Tenggara	430	8.0%	5.46	7.5	4	Moderate-High
Java	315	5.9%	5.43	7.0	1	Moderate

Gutenberg-richter analysis

At the zonal level, the b-value shows moderate variation that carries physical meaning. Sumatra returns a b-value near 0.95 and Western New Guinea near 0.97. Both values lie below the catalog average of 0.98 implying that these zones accumulate differential stress more efficiently and release a comparatively larger share of their seismic moment through infrequent but very large ruptures. In contrast, the Maluku-Sulawesi zone yields b approximately equal to 1.08 which consistent with a more fragmented stress field in which energy dissipates through frequent moderate earthquakes rather than through occasional great ones.

Parameter selection via k-distance plot

Figure 3a shows the k-distance plot that was constructed from the 5th-nearest-neighbor distances of all 5,364 events sorted in descending order. The curve descends steeply from the upper left before flattening into a long tail. The elbow identified by the maximum-curvature criterion that falls at a distance of 0.59 degrees or approximately 65 km. This value is consistent with the typical along-strike coherence length of seismogenic fault segments in the Indonesian arc system and with the resolution of BMKG hypocenter locations, which are generally better than 30 km for well-recorded events. Figure 3b illustrates how the choice of eps parameter affects the balance between cluster count and noise fraction at 0.3 degrees with 47 clusters emerge but noise reaches 34.8%. At 1.5 degrees, all events collapse into a single cluster as previously observed. The selected value of 0.59 degrees achieves 16 clusters with a

noise fraction of 8.6%. Figure 3c and 3d spatially illustrate all the events with and without noise, respectively.

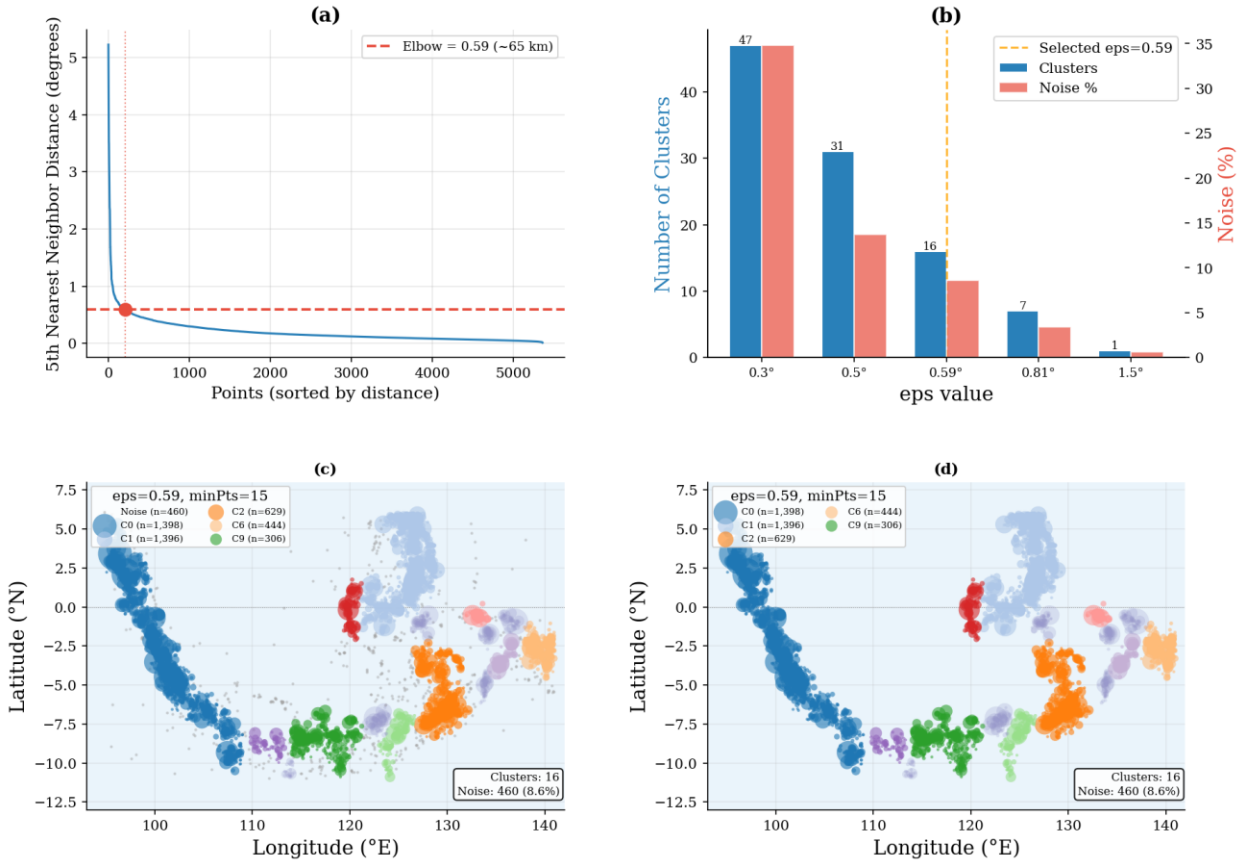

Figure 3. DBSCAN parameter selection and results. (a) k-distance plot with elbow marking the optimal eps = 0.59 degrees (about 65 km). (b) Sensitivity of cluster count and noise percentage to eps. (c) All events including noise points (grey). (d) Cluster events only with top-5 clusters labeled

Spatial structure of DBSCAN clusters

With eps parameter set to 0.59 degrees and minPts paremeter to 15, DBSCAN partitioned the catalog into 16 clusters and 460 noise points. The two largest clusters each contain roughly 1,400 events and correspond geographically to the Sumatra segment (Cluster 0, dominated by Sumatra zone events with Mmax = 9.0 from 17 major events) and the primary Maluku-Sulawesi segment (Cluster 1 with Mmax = 7.6 from 12 major events). A third sizable cluster (Cluster 2 with n = 629) captures a secondary Maluku-Sulawesi concentration, while Cluster 6 with n = 444 covers the Western New Guinea domain. Smaller clusters correspond to seismogenic segments along Bali-Nusa Tenggara and isolated zones in Western New Guinea and Sulawesi.

The 460 noise events warrant attention. Scattered at various locations away from the main arc system, they represent seismicity whose spatial density falls below the threshold required for cluster membership. Some may reflect catalog artifacts arising from poorly constrained hypocenters, particularly for events in remote offshore areas with sparse network coverage. Others may indicate the activation of intraplate structures or previously unmapped fault segments. Both possibilities merit targeted relocation studies using waveform cross-correlation which lies outside the scope of this paper but is identified as a priority for follow-up research.

Optimal partition and cluster characteristics

The K-means Elbow curve and Silhouette Scores shown in Figure 4b and 4c both point to K = 4 as the optimal partition. Inertia decreases rapidly from K = 2 to K = 4, after which additional clusters yield diminishing improvements. The Silhouette Score peaks at 0.386 for K = 4 and declines for larger values. The spatial distribution of the four clusters is mapped in Figure 4a.

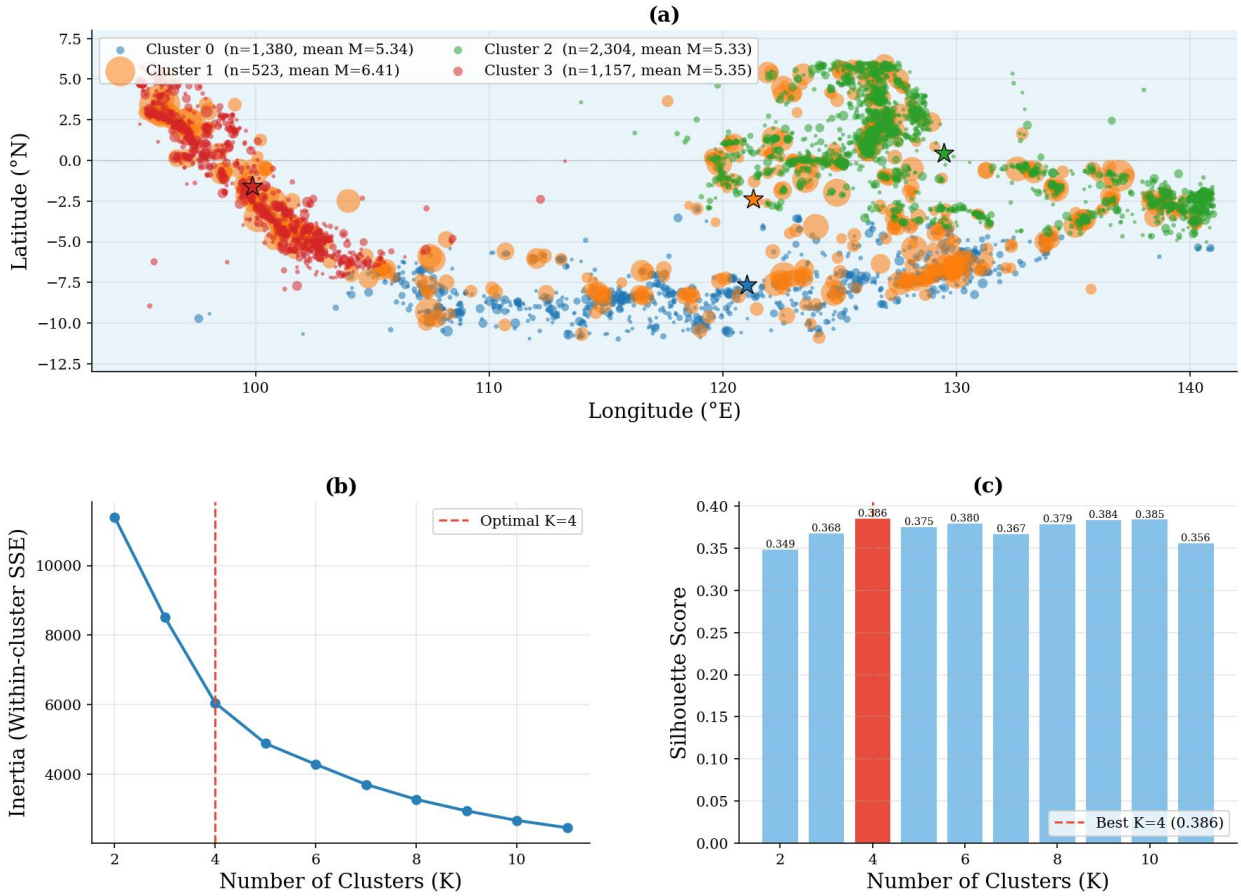

Figure 4. K-means clustering results (K = 4). (a) Spatial distribution with centroid positions marked by stars. (b) Elbow method showing within-cluster SSE vs K. (c) Silhouette Score vs K with the selected value highlighted

Table 2 summarizes the four clusters. Three of the four (Clusters 0, 2, and 3) exhibit nearly identical mean magnitudes of approximately 5.33-5.35 and maximum magnitudes no greater than M 6.2. They represent geographically differentiated expressions of background seismicity, corresponding broadly to the eastern arc, western arc, and Sumatra-centered environments, respectively. Cluster 1 is the outlier in every sense, although it contains only 523 events (9.8% of the catalog), it absorbs all 64 major earthquakes (M≥7.0) including the M 9.0 Sumatra event and an M 8.1 Western New Guinea rupture, and its mean magnitude of 6.41 exceeds the background clusters by more than one unit.

Table 2. K-means cluster statistics with K = 4

Cluster	Events (n)	Percentage Catalog	Mean M	Max M	$M \geq 7.0$	Seismic Regime
0	1,380	25.7%	5.34	6.0	0	Background seismicity

Cluster	Events (n)	Percentage Catalog	Mean M	Max M	$M \geq 7.0$	Seismic Regime
						(western arc)
1	523	9.8%	6.41	9.0	64	High-energy / major earthquake cluster
2	2,304	43.0%	5.33	6.0	0	Background seismicity (eastern arc)
3	1,157	21.6%	5.35	6.2	0	Sumatra-dominated background

Cluster profile analysis

Figure 5 displays the profile of each cluster across four diagnostic panels (a, b, c, and d). Figure 5a shows the event counts and its percentages of each cluster. The contrast between Cluster 1 and the background clusters is most apparent in panels Figure 5b and 5c with the mean magnitude gap of roughly 1.07 units corresponds to approximately twelve times the energy release per event under standard Richter scaling, and the maximum magnitude gap of three units represents a factor of roughly 31,000 in peak seismic moment. Figure 5d shows that Cluster 1 members are not confined to any single geographic zone. They appear in Sumatra, Maluku, Western New Guinea, and everywhere else along the arc, which is precisely what makes this cluster so significant from a hazard perspective.

The geographic correspondence between background clusters and conventional zones is close but not perfect. Cluster 2 whose centroid sits at approximately 129 degrees East aligns with the Maluku-Sulawesi domain. Cluster 3 centered near 100 degrees East maps onto the Sumatra zone. Cluster 0 with a centroid near 121 degrees East occupies a transitional position between the western and central arc. This correspondence validates the broad logic of the conventional zonation while simultaneously demonstrating that the high-energy cluster, which cuts across all zone boundaries, cannot be captured by any purely geographic scheme.

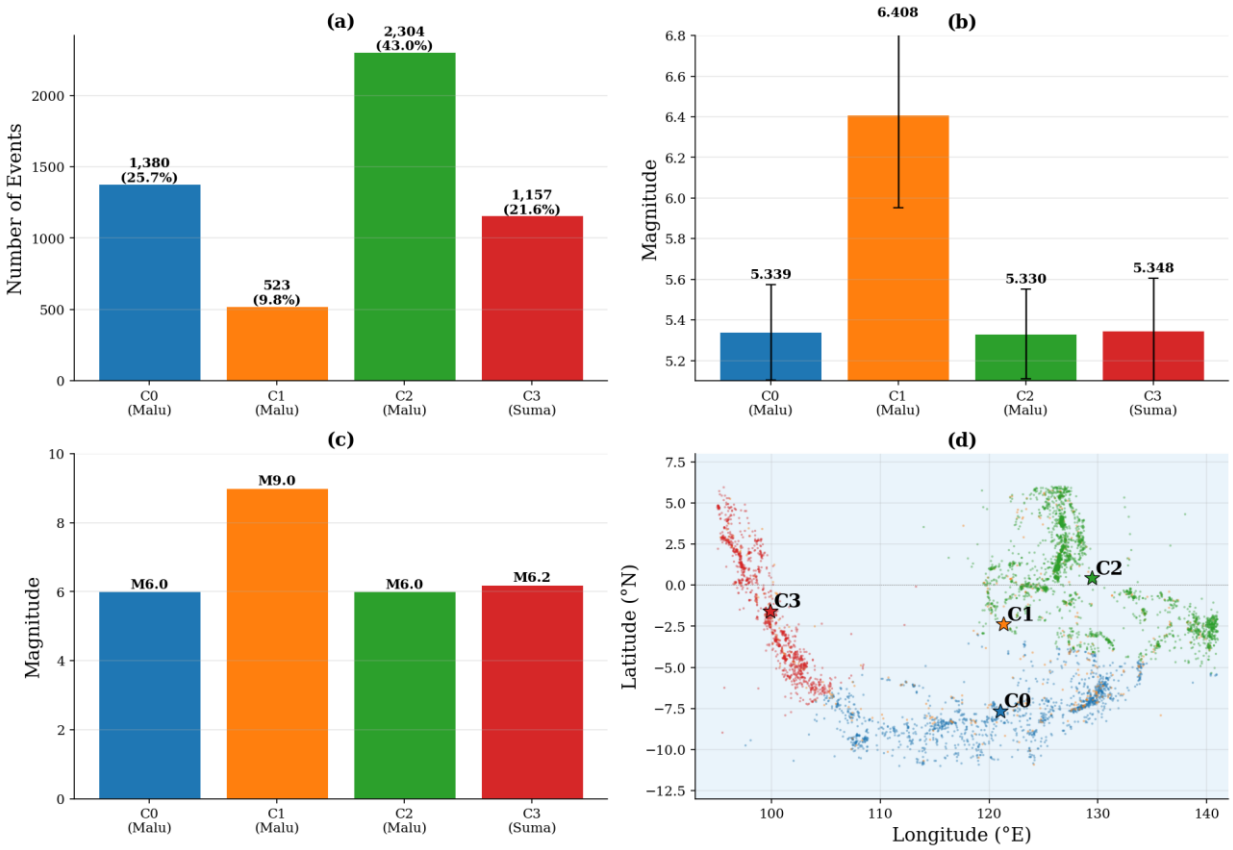

Figure 5. K-means cluster profile. (a) Event counts with percentage shares. (b) Mean magnitude with one standard deviation error bars. (c) Maximum recorded magnitude. (d) Centroid locations overlain on the base seismicity map

Discussion

DBSCAN and K-means answered different questions and produced complementary insights. DBSCAN, working in two-dimensional geographic space, resolved the structural question of how seismicity is spatially organized which 16 segments emerge at the 65 km scale, broadly corresponding to distinct portions of the subduction arc and back-arc fault systems. Its noise classification identified 460 events that fall outside any coherent seismogenic cluster.

K-means working in three-dimensional feature space that includes magnitude addressed the hazard question of how events are organized by energy character. The four-cluster solution revealed that high-energy events form a distinct population, Cluster 1, that is not localized to any single geographic zone. A zone-based hazard assessment would distribute these events across five provincial averages potentially underestimating the risk at specific sites that happen to lie on the pathways of great-rupture propagation. The K-means result provides a basis for targeting site-specific probabilistic analysis at locations where Cluster-1-type events are most likely to generate damaging ground motion.

The practical implication of Cluster 1 is straightforward that any infrastructure project, whether a hospital, a bridge, a dam, or a nuclear facility, that is sited within the geographic footprint of this high-energy cluster should be designed against ground motions compatible with M 7-9 events, regardless of which conventional zone the project nominally occupies. The conventional zone-average approach would assign Java, for example, a moderate hazard rating, yet Java falls within the spatial footprint of

Cluster 1, and the Mw 7.0 event recorded there during the study period confirms that the zone is not immune to major ruptures.

The 16 DBSCAN clusters also carry operational value for tsunami warning. Because each cluster corresponds to a spatially coherent seismogenic segment, a strong event in one cluster can be expected to transfer stress to adjacent clusters along the arc raising short-term rupture probabilities in neighboring segments. Early-warning systems that treat each segment as independent may underestimate the likelihood of triggered events. The DBSCAN segmentation provides a natural framework for segment-interaction modeling.

The results of this study align with and extend a growing body of literature applying machine learning to seismic catalog analysis. The accelerating adoption of ML in seismology reported by Beroza et al. (2021) and Mousavi and Beroza (2023) reflects a broader shift from expert-defined to data-driven approaches in earthquake source characterization. Sharma and Nanda (2022) showed that combining self-organizing maps with DBSCAN improves spatial zone coherence in earthquake-prone regions, while Piegari et al. (2024) demonstrated that hierarchical density clustering can illuminate fault segmentation at multiple scales simultaneously. The sensitivity of DBSCAN to its epsilon parameter was addressed in the present study through an objective k-distance plot analysis for reducing hyperparameter dependence (Sabermahani & Frederiksen, 2024). The eps value of 0.59 degrees obtained in this study is more conservative than the 1.5 degrees used in prior Indonesian applications, and the improvement in cluster count from one to sixteen confirms that parameter selection has first-order consequences for the interpretability of density-based seismogenic zonation.

The K-means results corroborate the finding of Yuan (2021) that incorporating magnitude as a third feature alongside geographic coordinates separates high-energy seismic populations from background seismicity more effectively than spatially exclusive partitioning. The present silhouette score of 0.386 is comparable to the 0.3245 reported by Dwitiyanti et al. (2024) for K-means applied to Indonesian BMKG data suggesting that the moderate cluster separation is a characteristic property of the Indonesian seismic dataset rather than a limitation of the algorithm or parameterization. The cross-zone distribution of Cluster 1 members echoes the seismogenic regime separations identified by Rehman et al. (2014) for Pakistan and by Katta et al. (2024) for Gujarat where K-means correctly distinguished mechanistically distinct earthquake populations that share geographic proximity but differ in energy character.

Overall, the DBSCAN segmentation and the K-means regime classification complement each other in precisely the manner that the broader seismological ML literature would predict such as density-based methods characterize the topological structure of the seismogenic system, while partition-based methods expose energy-based heterogeneity within it (Weatherill & Burton, 2009; Iaccarino & Picozzi, 2023). The Gutenberg-Richter analysis, implemented using the frequency-magnitude statistics framework established by Wiemer (2001), provides an independent physical validation for the cluster boundaries, since the zonal b-value differences are consistent with the stress-regime contrasts inferred from the K-means partition. This convergence of three independent analytical lines strengthens confidence in the zonation and suggests that the integrated approach introduced here is transferable to other complex arc systems where conventional tectonic boundary placement is contested or incomplete.

This research aligns with several United Nations Sustainable Development Goals. At the most direct level, the seismic hazard zonation produced here supports SDG 11 (Sustainable Cities and Communities) by supplying evidence-based spatial risk information that municipal governments, land-use planners, and building code committees across Indonesia can use to reduce earthquake exposure in rapidly urbanizing coastal settlements. The identification of a high-energy seismic cluster that crosses conventional tectonic zone boundaries reinforces the urgency of targeted infrastructure retrofitting and early-warning investment called for under SDG 13 (Climate Action and Disaster Risk Reduction). The methodological innovations introduced which objective eps selection via k-distance plot and multi-feature machine learning clustering contribute to SDG 9 (Industry, Innovation and Infrastructure) by offering a reproducible, data-driven alternative to subjective zonation that engineering practitioners can incorporate into probabilistic seismic hazard analyses (Khalqillah, et al., 2025). Finally, by documenting and openly sharing a transferable analytical workflow applicable to any seismically active region, this study advances SDG 17 (Partnerships for the Goals), supporting capacity building in disaster risk science across the broader Pacific and Indian Ocean seismic belts.

CONCLUSION

Fundamental Finding: A decade of Indonesian earthquake data (2016-2025, M >= 5.0, n = 5,364) was analyzed using conventional tectonic zone statistics, DBSCAN density-based clustering with an objectively determined eps parameter, and K-means partitioning optimized by Elbow and Silhouette criteria. Maluku-Sulawesi recorded the highest earthquake frequency at 40.8% of the catalog, yet Sumatra produced the most destructive seismicity, including an M 9.0 rupture, earning the highest composite hazard score; the catalog-wide Gutenberg-Richter b-value of 0.98 (R^2 = 0.996) is consistent with global subduction-zone averages, while lower zonal b-values in Sumatra and Western New Guinea (~0.95–0.97) point to elevated potential for infrequent but large-magnitude ruptures in those zones. An objective k-distance plot analysis identified an optimal DBSCAN eps of 0.59 degrees (~65 km), free of any prior assumption, and at this parameter DBSCAN resolved 16 spatially distinct seismogenic segments alongside 460 noise events (8.6%) that warrant targeted relocation studies. K-means clustering with K = 4 (Silhouette = 0.386) partitioned the catalog into four seismic regimes of which Cluster 1 (n = 523, mean M = 6.41) concentrated all 64 major earthquakes including events up to M 9.0, whereas the three background clusters (mean M ≈ 5.33–5.35) produced no event above M 6.2. **Implication:** DBSCAN defines the spatial segmentation of the seismogenic arc, while K-means reveals energy-based regimes that cut across conventional zone boundaries jointly exposing dimensions of hazard structure that geography-only zonation obscures and that have direct relevance for infrastructure design standards, land-use planning, and early-warning system configuration across the Indonesian archipelago. **Limitation:** The absence of focal depth and focal mechanism data in the catalog confines the analysis to two geographic dimensions, preventing discrimination between shallow megathrust, intermediate intraslab, and upper-crustal earthquake populations that differ fundamentally in their ground-motion characteristics and hazard footprint. The ten-year observation window, while sufficient for characterizing background seismicity, captures only a handful of great earthquakes and cannot reliably constrain recurrence intervals for M ≥ 8.0 ruptures. Furthermore, the moderate Silhouette Score of 0.386 for the K-means solution

reflects genuine spatial overlap between seismic populations along a continuous arc system, and the 460 noise events flagged by DBSCAN require targeted relocation studies before their tectonic significance can be evaluated. **Future Research:** Future work should address the limitations above by incorporating three-dimensional clustering extending the catalog into the pre-2016 instrumental period to better constrain recurrence intervals for great earthquakes, and couple the cluster-based zonation with ground-motion prediction equations and regional sub-catalog analyses to test whether locally optimized eps values uncover finer-scale seismogenic segmentation invisible at the national scale.

ACKNOWLEDGEMENTS

This research was supported by Penelitian Dasar Dana non APBN Universitas Negeri Surabaya research grant (343/UN38/HK/PP/2024).

AUTHOR CONTRIBUTIONS

Atik Wintarti contributed to the conceptualization of the study, research design, methodology development, supervision, and manuscript review. **Fadhilah Qalbi Annisa** contributed to data collection, data preprocessing, machine learning implementation, formal analysis, visualization, and manuscript drafting. **Harmon Prayogi** contributed to the conceptual framework, methodology development, validation, supervision, manuscript review and editing, and project administration. **Yuliani Puji Astuti** contributed to data interpretation, validation, visualization, manuscript review and editing, and supervision. **Ibnu Febry Kurniawan** contributed to methodology refinement, software implementation, formal analysis, validation, manuscript review and editing, and academic supervision. All authors have read, reviewed, and approved the final version of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors state that no financial or personal conflicts of interest exist that may have affected the content or findings of this research.

STATEMENT ON THE USE OF AI OR DIGITAL TOOLS IN WRITING

The authors acknowledge the use of digital tools, including AI-based technologies, during the preparation of this manuscript. ChatGPT (OpenAI) was used to assist with language refinement, grammar correction, academic writing improvement, and manuscript organization. Python (Scikit-learn) was employed for data preprocessing, machine learning implementation (DBSCAN and K-means clustering), and analytical computations. All AI-assisted outputs and digital analyses were carefully reviewed, verified, and revised by the authors to ensure the accuracy, originality, and integrity of the research. The authors take full responsibility for the content and conclusions presented in this manuscript.

REFERENCES

Abers, G. A., & McCaffrey, R. (1988). Active deformation in the New Guinea fold-and-thrust belt: Seismological evidence for strike-slip faulting and basement-involved thrusting. *Journal of Geophysical Research: Solid Earth, 93*(B11), 13332–13354. https://doi.org/10.1029/JB093iB11p13332

Beroza, G. C., Segou, M., & Mostafa Mousavi, S. (2021). Machine learning and earthquake forecasting: Next steps. *Nature Communications, 12,* 4761. https://doi.org/10.1038/s41467-021-24952-6

Bilek, S. L., & Lay, T. (2018). Subduction zone megathrust earthquakes. *Geosphere, 14*(4), 1468–1500. https://doi.org/10.1130/GES01608.1

Cilia, M. G., Mooney, W. D., & Nugroho, C. (2022). Field insights and analysis of the 2018 Mw 7.5 Palu, Indonesia earthquake, tsunami and landslides. Pure and *Applied Geophysics, 179*, 1291–1323. https://doi.org/10.1007/s00024-021-02852-6

Curray, J. R. (2005). Tectonics and history of the Andaman Sea region. *Journal of Asian Earth Sciences, 25*(1), 187–232. https://doi.org/10.1016/j.jseaes.2004.09.001

Damanik, R., Gunawan, E., Widiyantoro, S., Supendi, P., Atmaja, F.W., Ardianto, Husni, Y.M.R., Zulfakriza and Sahara, D.P. (2023). New assessment of the probabilistic seismic hazard analysis for the greater Jakarta area, Indonesia. *Geomatics, Natural Hazards and Risk*, *14*(1), 2202805.

Dwitiyanti, N., Kumala, S. A., & Handayani, S. D. (2024). Comparative study of earthquake clustering in Indonesia using K-Medoids, K-Means, DBSCAN, Fuzzy C-Means and K-AP algorithms. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 8*(6), 768–778. https://doi.org/10.29207/resti.v8i6.5514

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

Heidarzadeh, M., & Mulia, I. E. (2022). A new dual earthquake and submarine landslide source model for the 28 September 2018 Palu (Sulawesi), Indonesia tsunami. *Coastal Engineering Journal, 65*(1), 1–20. https://doi.org/10.1080/21664250.2022.2122293

Hutchings, S. J., & Mooney, W. D. (2021). The seismicity of Indonesia and tectonic implications. *Geochemistry, Geophysics, Geosystems, 22*(9), e2021GC009812. https://doi.org/10.1029/2021GC009812

Iaccarino, A. G., & Picozzi, M. (2023). Detecting the preparatory phase of induced earthquakes at the Geysers (California) using K-means clustering. *Journal of Geophysical Research: Solid Earth*, *128*(10), e2023JB026429. https://doi.org/10.1029/2023JB026429

Irsyam, M., Widiyantoro, S., Natawidjaja, D. H., Meilano, I., Rudiyanto, A., Hidayati, S., Triyoso, W., Hanifa, N. R., Djarwadi, D., Fadhli, L., & Sunarjito. (2020). Development of the 2017 national seismic hazard maps of Indonesia. *Earthquake Spectra, 36*(S1), 112–136. https://doi.org/10.1177/8755293020951206

Iuchi, K., Takagi, H., Jibiki, Y., Kondo, T., Kusunoki, A., Hanifa, N. R., Pelupessy, D., Gayathri, R. T., & Olshansky, R. (2023). Questioning the hazard map-based rebuilding process: learning from the 2018 Sulawesi earthquake in Indonesia. *Coastal Engineering Journal, 65*(1), 126–148. https://doi.org/10.1080/21664250.2023.2165430

Jena, R., Pradhan, B., Beydoun, G., Al-Amri, A., & Sofyan, H. (2020). Earthquake hazard and risk assessment using machine learning approaches at Palu, Indonesia. *Science of The Total Environment, 749*, 141582. https://doi.org/10.1016/j.scitotenv.2020.141582

Jihad, A., Muksin, U., Syamsidik, & Ramli, M. (2021). Earthquake relocation to understand the megathrust segments along the Sumatran subduction zone. *IOP*

Conference Series: Earth and Environmental Science, 630, 012002. https://doi.org/10.1088/1755-1315/630/1/012002

Kadir, A., Sukri, A. S., Aswad, N. H., Masdiana, & Nasrul. (2024). Evolution and implications of changes in seismic load codes for earthquake resistant structures design. *Civil Engineering Journal, 10*(1), 62–82. https://doi.org/10.28991/CEJ-2024-010-01-04

Katta, V. S., Shrivastava, G., & Kushwah, V. K. (2024). Hierarchical K-means clustering of earthquake data for in-depth spatial and magnitude analysis in Gujarat, India. *Multidisciplinary Science Journal, 7*(7), 2025325. https://doi.org/10.31893/multiscience.2025325

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281–297. https://cir.nii.ac.jp/crid/1571135649659368064

Marliyani, G. I., Arrowsmith, J. R., & Whipple, K. X. (2016). Characterization of slow slip rate faults in humid areas: Cimandiri Fault Zone, Indonesia. *Journal of Geophysical Research: Earth Surface, 121*(12), 2287–2308. https://doi.org/10.1002/2016JF003846

McCaffrey, R. (1982). Lithospheric deformation within the Molucca Sea arc-arc collision: Evidence from shallow and intermediate earthquake activity. *Journal of Geophysical Research: Solid Earth, 87*(B5), 3663–3678. https://doi.org/10.1029/JB087iB05p03663

McCaffrey, R. (2009). The tectonic framework of the Sumatran subduction zone. *Annual Review of Earth and Planetary Sciences, 37*, 345–366. https://doi.org/10.1146/annurev.earth.031208.100212

McNeill, L. C., & Henstock, T. J. (2014). Forearc structure and morphology along the Sumatra-Andaman subduction zone. *Tectonics, 33*(2), 112–134. https://doi.org/10.1002/2012TC003264

Mousavi, S. M., & Beroza, G. C. (2023). Machine learning in earthquake seismology. *Annual Review of Earth and Planetary Sciences, 51*, 105–129. https://doi.org/10.1146/annurev-earth-071822-100323

Natawidjaja, D. H., & Triyoso, W. (2007). The Sumatran fault zone: from source to hazard. *Journal of Earthquake and Tsunami, 1*(01), 21–47. https://doi.org/10.1142/S1793431107000031

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Piegari, E., Camanni, G., Mercurio, M., & Marzocchi, W. (2024). Illuminating the hierarchical segmentation of faults through an unsupervised learning approach applied to clouds of earthquake hypocenters. *Earth and Space Science, 11*(10), e2023EA003267. https://doi.org/10.1029/2023EA003267

Khalqillah, A., Umar, M., Simanjuntak, A. V., Jihad, A., & Banyunegoro, V. H. (2025). Seismic hazard estimation for Sumatra and Kalimantan region using event-based probabilistic seismic hazard analysis (EB-PSHA). *Journal of Geoscience, Engineering, Environment, and Technology*, *10*(3), 329-337. https://doi.org/10.25299/jgeet.2025.10.3.21936

Pusat Studi Gempa Nasional (PusGen). (2017). Peta Sumber dan Bahaya Gempa Indonesia Tahun 2017. Kementerian PUPR.

R, S. R., & Madrinovella, I. (2024). Spatial and temporal b-value analysis of the Yogyakarta region using earthquake data 1960–2024. *JGE (Jurnal Geofisika Eksplorasi), 10*(3), 191–203. https://doi.org/10.23960/jge.v10i3.468

Rehman, K., Burton, P. W., & Weatherill, G. A. (2014). K-means cluster analysis and seismicity partitioning for Pakistan. *Journal of Seismology, 18*(3), 401–419. https://doi.org/10.1007/s10950-013-9415-y

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

Sabermahani, S., & Frederiksen, A. W. (2024). Improved earthquake clustering using a density-adaptive DBSCAN algorithm: An example from Iran. *Seismological Research Letters, 95*(2A), 942–951. https://doi.org/10.1785/0220220305

Sari, E. P., Kusmita, T., & Kurniawan, W. B. (2025). The Temporal Variation b-Value Temporal Variation of b-Value in Bengkulu Province Using the Maximum Likelihood Method (Earthquake Case Study 2012-2022). *Jurnal Riset Fisika Indonesia*, *6*(01), 30-39.

Sarkawi, G. M., Feng, L., McCaughey, J. W., Meltzner, A. J., Susilo, S., Muksin, U., Socquet, A., Oktari, R. S., Adi, S. P., Burgmann, R., & Hill, E. M. (2024). Insights into tectonic hazards since the 2004 Indian Ocean earthquake and tsunami. *Nature Reviews Earth and Environment, 5*, 791–808. https://doi.org/10.1038/s43017-024-00613-2

Schorlemmer, D., Wiemer, S., & Wyss, M. (2005). Variations in earthquake-size distribution across different stress regimes. *Nature, 437*(7058), 539–542. https://doi.org/10.1038/nature04094

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

Sharma, V. K., Vijay, R. K., & Nanda, S. J. (2022). Identification and spatio-temporal analysis of earthquake clusters using SOM-DBSCAN model. *Neural Computing and Applications, 35*, 4211–4230. https://doi.org/10.1007/s00521-022-08085-5

Socquet, A., Simons, W., Vigny, C., McCaffrey, R., Subarya, C., Sarsito, D., Ambrosius, B., & Spakman, W. (2006). Microblock rotations and fault coupling in SE Asia triple junction (Sulawesi, Indonesia) from GPS and earthquake slip vector data. *Journal of Geophysical Research: Solid Earth, 111*(B8), B08409. https://doi.org/10.1029/2005JB003963

Stein, S., & Wysession, M. (2003). *An Introduction to Seismology, Earthquakes, and Earth Structure*. Blackwell Publishing.

Supendi, P., Nugraha, A. D., Widiyantoro, S., Abdullah, C. I., Rawlinson, N., Cummins, P. R., Daryono, D., Wiyono, S. H., Shiddiqi, H. A., & Rosalia, S. (2020). Relocated aftershocks and background seismicity in eastern Indonesia shed light on the 2018 Lombok and Palu earthquake sequences. *Geophysical Journal International, 221*(3), 1845–1855. https://doi.org/10.1093/gji/ggaa118

Tregoning, P., Brunner, F. K., Bock, Y., Puntodewo, S. S. O., McCaffrey, R., Genrich, J. F., Calais, E., Rais, J., & Subarya, C. (1994). First geodetic measurement of convergence across the Java Trench. *Geophysical Research Letters, 21*(19), 2135–2138. https://doi.org/10.1029/94GL01856

Vijay, R. K., & Nanda, S. J. (2023). Earthquake pattern analysis using subsequence time series clustering. *Pattern Analysis and Applications, 26*(1), 19–37. https://doi.org/10.1007/s10044-022-01092-1

Wajedy, M. F., Massinai, M. A., Fahruddin, F., & Thariq, A. (2026). Seismic hazard and tectonic stress in Halmahera, Indonesia based on b-value and apparent stress analyses. *Physics of the Earth and Planetary Interiors*, 107512. https://doi.org/10.1016/j.pepi.2026.107512

Weatherill, G., & Burton, P. W. (2009). Delineation of shallow seismic source zones using K-means cluster analysis, with application to the Aegean region. *Geophysical Journal International, 176*(2), 565–588. https://doi.org/10.1111/j.1365-246X.2008.03997.x

Widiyantoro, S., & van der Hilst, R. (1997). Mantle structure beneath Indonesia inferred from high-resolution tomographic imaging. *Geophysical Journal International, 130*(1), 167–182. https://doi.org/10.1111/j.1365-246X.1997.tb00996.x

Widiyantoro, S., Supendi, P., Ardianto, A., Baskara, A. W., Bacon, C. A., Damanik, R., Rawlinson, N., Gunawan, E., Sahara, D. P., Zulfakriza, Z., & Mori, J. (2022). Implications for fault locking south of Jakarta from an investigation of seismic activity along the Baribis fault, northwestern Java, Indonesia. *Scientific Reports, 12*(1), 10143. https://doi.org/10.1038/s41598-022-13896-6

Wiemer, S. (2001). A software package to analyze seismicity: ZMAP. *Seismological Research Letters, 72*(3), 373–382. https://doi.org/10.1785/gssrl.72.3.373

Wulandari, R., Chan, C. H., & Wibowo, A. (2023). The 2022 Mw6.2 Pasaman, Indonesia, earthquake sequence and its implication of seismic hazard in central-west Sumatra. *Geoscience Letters, 10*, 7. https://doi.org/10.1186/s40562-023-00279-6

Yuan, R. (2021). An improved K-means clustering algorithm for global earthquake catalogs and earthquake magnitude prediction. *Journal of Seismology, 25*(3), 1005–1020. https://doi.org/10.1007/s10950-021-09999-8

Zhang, Q., Guo, F., Zhao, L., & Wu, Y. (2017). Geodynamics of divergent double subduction: 3-D numerical modeling of a Cenozoic example in the Molucca Sea region, Indonesia. *Journal of Geophysical Research: Solid Earth, 122*(5), 3977–3998. https://doi.org/10.1002/2017JB013991

Zulfakriza, Z., Nugraha, A. D., Heryandoko, N., Ry, R. V., Muttaqy, F., Andika, A., Azhari, M. F., Putra, A. S., Palgunadi, K. H., Cummins, P. R., & Supendi, P. (2024). Seismic source analysis of the destructive Mw 5.6 Cianjur (Indonesia) earthquake from relocated aftershocks. *Natural Hazards and Earth System Sciences, 24*, 261–274. https://doi.org/10.1038/s41598-024-60408-9

Atik Wintarti
Department of Data Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, 60231, Surabaya, Indonesia
Building E2, 2nd Floor, Unesa Campus, Jl. Ketintang, Ketintang, Gayungan District, Surabaya, East Java 60231
Email: atikwintarti@unesa.ac.id

Fadhilah Qalbi Annisa
Department of Artificial Intelligence, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, 60231, Surabaya, Indonesia
Building E2, 2nd Floor, Unesa Campus, Jl. Ketintang, Ketintang, Gayungan District, Surabaya, East Java 60231
Email: fadhilahannisa@unesa.ac.id

*** Harmon Prayogi (Corresponding Author)**
Department of Artificial Intelligence, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, 60231, Surabaya, Indonesia
Building E2, 2nd Floor, Unesa Campus, Jl. Ketintang, Ketintang, Gayungan District, Surabaya, East Java 60231
Email: harmonprayogi@unesa.ac.id

Yuliani Puji Astuti
Department of Data Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, 60231, Surabaya, Indonesia
Building E2, 2nd Floor, Unesa Campus, Jl. Ketintang, Ketintang, Gayungan District, Surabaya, East Java 60231
Email: yulianipuji@unesa.ac.id

Ibnu Febry Kurniawan
Department of Data Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, 60231, Surabaya, Indonesia;
Innovation Center for Artificial Intelligence, Universitas Negeri Surabaya, 60213 Surabaya, Indonesia
Building E2, 2nd Floor, Unesa Campus, Jl. Ketintang, Ketintang, Gayungan District, Surabaya, East Java 60231
Email: ibnufebry@unesa.ac.id