e-ISSN : 2722-4368

Spatiotemporal Dynamics of Seismic Activity in the Toba Caldera based on DBSCAN Clustering Algorithms

Marzuki Sinambela[1*], Purwantiningsih[2], Febria Anita[3], Puji Hartoyo[4], Ari Mutanto[5]
[12345]Program Studi Fisika, Fakultas Teknik dan Sains, Universitas Nasional, Jakarta
*e-mail penulis korespondensi: sinambela.m@gmail.com, purwantiningsih.fisika@gmail.com, febria.anita85@gmail.com, phartoyo310@gmail.com, arimutanto.mutanto696@gmail.com

Abstrak. *One of the most important volcanic systems in the world, the Toba Caldera is distinguished by a complicated seismotectonic environment brought about by interactions between regional tectonics, volcanic processes, and caldera dynamics. The performance of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is the main subject of this work, which uses unsupervised machine learning approaches to evaluate seismicity within the Toba Geopark between 2019 and 2022. As input features, earthquake parameters like as magnitude, peak ground acceleration, hypocentral depth, and geographic position were employed. The Davies-Bouldin index, the Calinski-Harabasz index, and the silhouette coefficient were used to assess clustering performance. In contrast to performance observed in tectonically influenced locations, the results show that DBSCAN is highly effective in volcanic environments, reaching a silhouette score of 0.679 and a Davies–Bouldin index of 0.404. The discrete and structure-controlled nature of volcanic seismicity is reflected in DBSCAN's identification of six compact seismic clusters linked to unique caldera-related structures and its classification of 91.48% of events as noise. Strong correlations between the detected clusters and recognized volcanic characteristics, such as caldera rims, central volcanic complexes, and interacting fault systems, are revealed by spatial analysis. These results demonstrate the environment-specific efficacy of clustering algorithms and aid in the creation of machine learning-based methods for hazard assessment and volcano seismic monitoring.*

Kata Kunci : *DBSCAN Clustering Algorithms, seismicity activity, Caldera Toba*

Abstract. *Salah satu sistem vulkanik terpenting di dunia, Kaldera Toba, memiliki lingkungan seismotektonik yang kompleks akibat interaksi antara tektonik regional, proses vulkanik, dan dinamika kaldera. Kinerja algoritma Density-Based Spatial Clustering of Applications with Noise (DBSCAN) menjadi fokus utama penelitian ini, yang menggunakan pendekatan pembelajaran mesin tanpa pengawasan untuk mengevaluasi aktivitas seismik di Toba Geopark antara tahun 2019 dan 2022. Sebagai fitur masukan, parameter gempa seperti magnitudo, percepatan tanah maksimum, kedalaman hiposentrum, dan posisi geografis digunakan. Indeks Davies-Bouldin, indeks Calinski-Harabasz, dan koefisien siluet digunakan untuk menilai kinerja pengelompokan. Berbeda dengan kinerja yang diamati di lokasi yang dipengaruhi tektonik, hasil menunjukkan bahwa DBSCAN sangat efektif di lingkungan vulkanik, mencapai skor siluet 0,679 dan indeks Davies-Bouldin 0,404. Sifat diskrit dan terkendali struktur dari seismisitas vulkanik tercermin dalam identifikasi DBSCAN terhadap enam kluster seismik kompak yang terkait dengan struktur kaldera unik dan klasifikasinya terhadap 91,48% peristiwa sebagai noise. Korelasi yang kuat antara kluster yang terdeteksi dan karakteristik vulkanik yang dikenal, seperti tepi kaldera, kompleks vulkanik pusat, dan sistem patahan yang berinteraksi, terungkap melalui analisis spasial. Hasil ini menunjukkan keefektifan algoritma klustering yang spesifik lingkungan dan membantu dalam pengembangan metode berbasis pembelajaran mesin untuk penilaian risiko dan pemantauan seismik.*

Keywords : *Algoritma Pengelompokan DBSCAN, aktivitas seismik, Kaldera Toba*

INTRODUCTION

The Toba Caldera is still an active volcanic-tectonic system despite its long history. Seismic activity within and around the caldera is still influenced by hydrothermal circulation, fault reactivation, and ongoing crustal deformation. The Sumatran fault system's broad tectonic stresses as well as localized activities including magma migration, pressure variations in hydrothermal

reservoirs, and structural modifications along caldera ring faults are all reflected in the earthquakes that have been recorded in this area. Consequently, the Toba region's seismicity displays a wide range of waveform features and significant geographical diversity. Analytical techniques that can distinguish and segregate signals originating from various sources while maintaining their spatial linkages within the caldera structure are necessary for accurately characterizing these seismic patterns. For such a complicated situation, traditional one-dimensional or solely temporal approaches are frequently inadequate. Rather, techniques that integrate various seismic characteristics and take into account three-dimensional subsurface geometry are crucial for uncovering the fundamental mechanisms that control caldera growth. By using such sophisticated analytical frameworks, scientists can learn more about the dynamic behavior of volcanic calderas like Toba, enhancing our comprehension of their historical development and possible future activity [1].

This complexity is especially noticeable in the Toba region. Volcanic-tectonic processes connected to caldera ring faults, resurgent structures, and near-surface deformation are frequently linked to shallow seismic occurrences. Seismicity at intermediate depths may be a reflection of magmatic processes in the crustal magma plumbing system, such as cooling, pressurization, or melt migration. In contrast, regional tectonics—particularly the large-scale deformation linked to the Sumatran fault system and subduction-related stresses—have a greater influence on deeper earthquakes. Beneath the caldera, these depth-dependent processes combine to form a complex seismic architecture that is both laterally and vertically oriented. Seismicity's multi-scale and multi-process nature presents substantial analytical difficulties. The fundamental causes of observed earthquake patterns may not be sufficiently resolved by conventional methods that depend on oversimplified assumptions of source mechanisms or homogeneous media. Because of this, there is growing interest in using sophisticated computational techniques that can handle high-dimensional data and capture minute changes in seismic behavior. These techniques have the potential to expose hidden structural controls, separate overlapping signals, and offer a more thorough knowledge of the dynamic processes forming volcanic systems such as the Toba Caldera [2]–[6].

In order to gain a better understanding of the spatial organization of seismicity within the Toba Caldera system, this study investigates the use of computer clustering approaches. Volcanic areas like Toba are ideal for data-driven analytical techniques because they produce dense and irregular seismic patterns that frequently defy straightforward classification [7]. This study aims to evaluate the capacity of three popular clustering algorithms—DBSCAN in particular—to catch significant seismic patterns in a complex volcanic setting. Both quantitative performance and geological interpretability are evaluated through a methodical comparison of algorithm outputs utilizing several validation metrics. The resulting clusters are analyzed in terms of their spatial coherence and connection to recognized fault systems, caldera structures, and depth-dependent seismic processes, in addition to numerical measurements. This dual viewpoint guarantees that the clustering results are both physically significant in a volcanological context and statistically sound. By using this method, the study seeks to determine which clustering techniques work best for examining volcanic seismicity that exhibits significant heterogeneity and multi-scale behavior. Additionally, by showing how sophisticated computational techniques can improve the understanding of intricate earthquake datasets, the results provide a broader contribution to the science of seismology. This work contributes to the creation of more trustworthy analytical frameworks for tracking and comprehending active volcanic systems by connecting algorithmic performance with geological insight.

METHODOLOGY

The Toba Caldera complex in North Sumatra, Indonesia, was chosen as the study's focus because it combines a well-defined volcanic architecture with a wealth of seismic monitoring data. The crater, which is around 100 km long and 30 km wide, is encircled by elevated volcanic topography that documents a lengthy history of eruptive and structural evolution. Lake Toba is located

at the center of the caldera. Toba is one of the most notable caldera systems on Earth due to the size and preservation of these features. Because the subsurface structure of this area represents the interaction of several geological processes acting at various depths, it is especially well adapted for the research of volcanic seismicity [1], [8], [9]. A complex three-dimensional framework is created by the coexistence of caldera ring faults, resurgent domes, and related fracture systems with wider tectonic impacts from the regional fault network. Within this paradigm, seismic activity records signals from tectonic, hydrothermal, and magmatic sources, providing rich data on ongoing subsurface dynamics. The Toba Caldera is a useful natural laboratory for researching earthquake clustering and spatial pattern creation in volcanic contexts because of its structural complexity and excellent seismic monitoring. In addition to providing insights into the processes that continue to shape one of the largest volcanic systems in the world, the study can investigate how computational tools react to real-world geological complexity by analyzing seismicity in this context.

The Meteorology, Climatology, and Geophysical Agency (BMKG) of Indonesia collected the seismic data utilized in this study from January 2019 to December 2022. This period of time offers a long enough observational window to record both episodic fluctuations associated with tectonic or volcanic activity as well as continuous background seismicity. The dataset includes comprehensive data for every earthquake, such as the event's magnitude, estimated peak ground acceleration, hypocentral depth, and geographic location determined by longitude and latitude. The dataset was subjected to a number of quality control processes meant to preserve internal consistency and completeness in order to guarantee the dependability of ensuing studies. Spatial filtering was used to keep only earthquakes that occurred within the Toba Caldera and its near surroundings, and events with missing or unclear parameters were eliminated. This method preserves signals pertinent to caldera-scale dynamics while reducing the impact of unrelated regional seismicity.

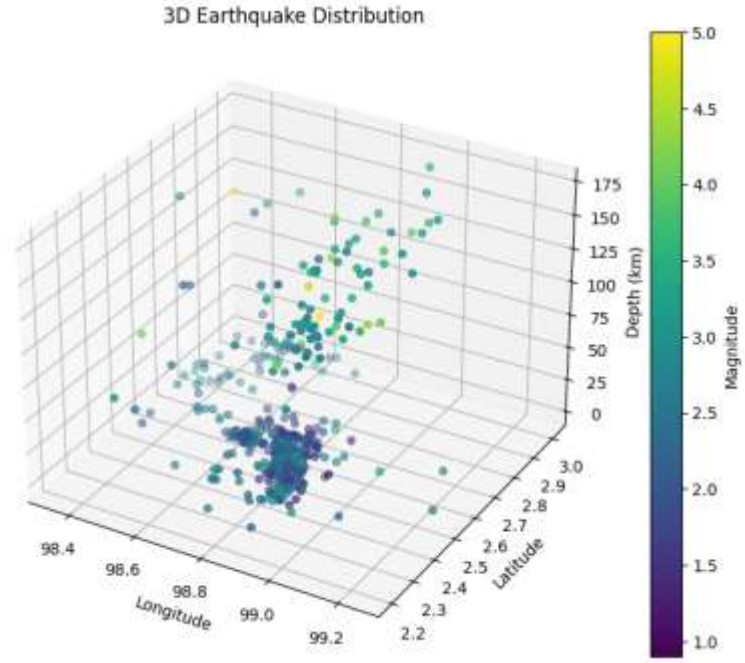

Figure 1. 3 D Earthquake Distribution in Toba

An unsupervised clustering approach called Density-Based Spatial Clustering of Applications with Noise (DBSCAN) groups data points based on their spatial density rather than how far they are from a predetermined cluster center. DBSCAN's fundamental concept is that clusters are areas in the data space with densely packed points, divided by areas with lower point densities. Because of this, the approach is especially well-suited for examining intricate, asymmetrical patterns that frequently

occur in natural systems. DBSCAN depends on two important factors. The first is the neighborhood radius, also known as epsilon (ε), which specifies the greatest distance between two places. The minimal number of points (MinPts) needed to create a dense region is the second parameter. If a data point has at least MinPts neighbors in its ε-neighborhood, it is considered a core point. Points that are not part of any dense region are referred to as noise, whereas points that are close to a core point but do not meet the density criteria are called border points. DBSCAN expands clusters through adjacent high-density locations by joining core sites that are density-reachable from one another. Because of this process, the algorithm can identify clusters of any shape without needing to know how many clusters there are. Because it can simultaneously filter out scattered background seismicity that does not reflect coherent physical processes and identify localized zones of elevated earthquake density associated with particular geological structures, such as caldera ring faults or magma pathways, DBSCAN is especially useful in volcanic seismicity studies.

To guarantee that the clustering results appropriately reflect the underlying seismic patterns, the algorithm parameters were meticulously tuned through an organized review procedure. For DBSCAN, k-distance analysis, which looks at the distribution of distances to each point's k-th nearest neighbor to determine a distinctive scale of local density, influenced the choice of the neighborhood radius (ε) and the minimum number of points needed to form a cluster. This method aids in differentiating between noise that is poorly distributed and densely concentrated areas. To ensure that the parameters are well suited to the spatial scales of volcanic features such caldera faults, resurgent domes, and magma conduits, neighborhood density calculations were also carried out to account for changes in seismic point distribution across the research area. The clustering technique can provide a reliable depiction of seismically active zones within the Toba Caldera system by methodically adjusting these parameters, which minimizes the misclassification of single events and efficiently captures meaningful clusters of earthquakes.

DBSCAN starts by selecting a random point and checking its ε-neighborhood. If the point is a core point ($|N\varepsilon(p)| \geq MinPts$), then the algorithm forms a new cluster and adds all density-reachable points to that cluster. This process is repeated for new points until no more points can be added. If the selected point is not a core point, it is categorized as a border point or noise depending on its relationship with other core points [10]–[14].

The ε-neighborhood of point p:

$$N\varepsilon(p) = \{q \mid d(p,q) \leq \varepsilon\}$$

Point p is a core point if:

$$|N\varepsilon(p)| \geq MinPts$$

Point q is density-reachable from p if:

$$q \in N\varepsilon(p)$$

and p is a core point.

The advantage of DBSCAN is that it can form arbitrary clusters and ignore noise. However, DBSCAN depends on the selection of appropriate parameters ε and MinPts. The worst-case time complexity of DBSCAN is $O(n^2)$, making it less efficient for large datasets. DBSCAN is also less effective if the density of clusters varies greatly [15]–[19].

RESULTS AND ANALYSIS

The Toba Caldera's core region, about at 98.8°E longitude and 2.5°N latitude, is where the majority of seismic activity is concentrated. This region corresponds to the caldera's main structural depression and the volcanic complexes that surround it. The distribution of seismic events in this zone is essentially circular, reflecting the caldera's general geometry and demonstrating the powerful structural control that the collapsed volcanic system exercised. Apart from the primary cluster, localized structural and magmatic impacts are reflected in secondary zones of enhanced seismicity at

periphery volcanic centers and along the caldera rim. Large stretches of relatively low seismicity separate these clearly defined high-activity zones. Density-based clustering techniques work best in such a geographical pattern, which is defined by discrete clusters embedded inside sparsely inhabited areas. Algorithms such as DBSCAN can efficiently detect significant seismic clusters that correlate to geological characteristics within the caldera system by utilizing the contrast between densely active and relatively inactive areas.

Using the volcanic seismicity dataset, a quantitative assessment of clustering performance reveals significant variations across the three techniques. Specifically, DBSCAN shows a great ability to find significant geographical groups in the data. The main areas of increased seismic activity were captured by the program, which identified six different clusters. Its efficacy is further supported by performance metrics: the Davies-Bouldin index of 0.404 confirms low cluster similarity, the Calinski-Harabasz index of 120.9 reflects high intra-cluster cohesion and inter-cluster separation, and the silhouette score of 0.679 shows that points are well matched to their assigned clusters while remaining distinct from nearby clusters. In this situation, DBSCAN's handling of sparse background activity is a crucial aspect. The method successfully eliminates dispersed seismic events that do not match cohesive geological structures, highlighting clusters connected to the major caldera and surrounding volcanic features, with a noise ratio of 91.48%. These findings show that DBSCAN, which offers both strong cluster identification and distinct separation of noise from structurally meaningful patterns, is especially well-suited for assessing heterogeneous volcanic seismicity.

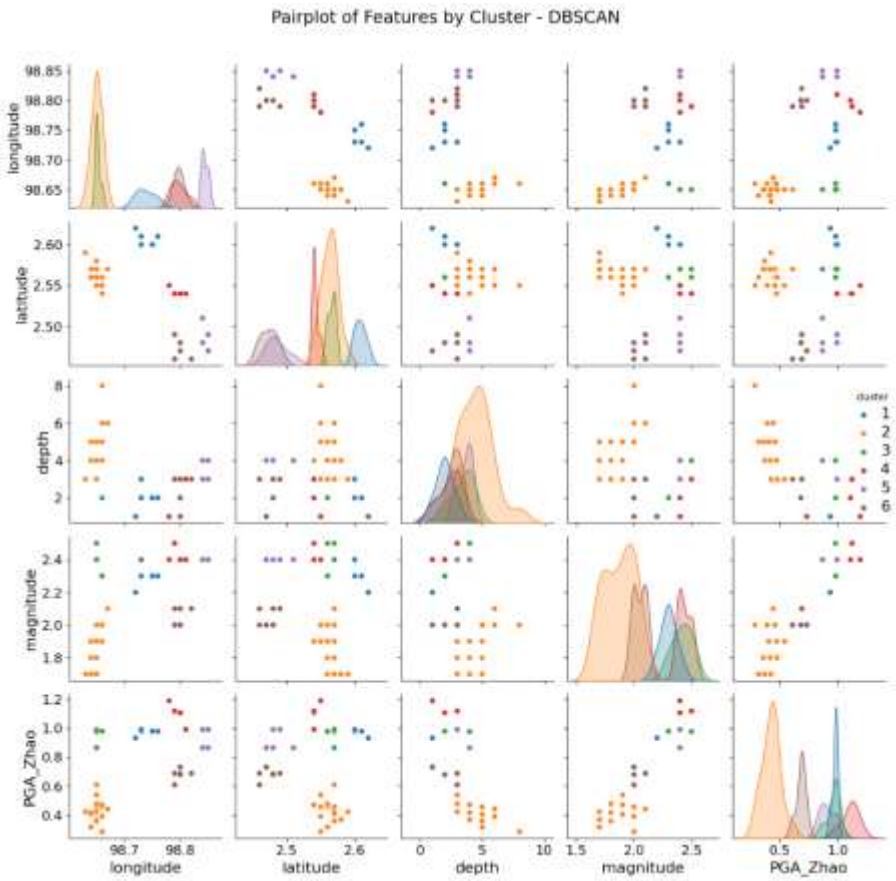

Figure 2. Pairplot of Fetaures by Cluster DBSCAN

When tested using a variety of clustering validation criteria, DBSCAN performs exceptionally well, especially in the Davies-Bouldin index and silhouette coefficient. These findings demonstrate the algorithm's ability to differentiate large amounts of background activity as noise while simultaneously recognizing compact, well-separated clusters. Because background events are widely

dispersed and active regions are frequently spatially confined around certain geological formations, such behavior is particularly well adapted to volcanic seismicity. Even though 91.48% of events in this study are classified as noise by the algorithm, this result is not a restriction but rather an accurate representation of the underlying seismicity patterns. Most earthquakes in volcanic environments, such as the Toba Caldera, are isolated events or extremely small, scattered groups that do not form significant clusters around large buildings. In contrast, seismicity tends to produce more continuous or linear patterns along faults in regional tectonic contexts. DBSCAN provides a more accurate and geologically significant identification of clustered seismic zones by efficiently removing these scattered events, closely aligning the clustering results with the physical reality of volcanic processes.

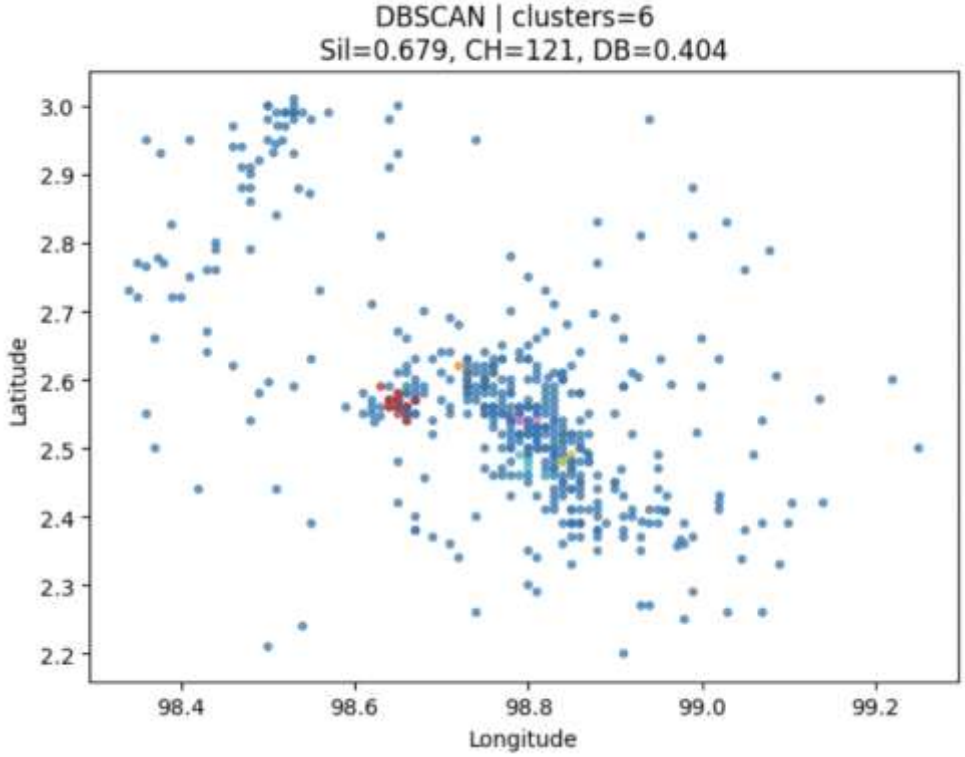

Figure 3. DBSCAN Cluster visualizations

The high silhouette coefficient seen is supported by the six-cluster solution, which shows good separation across all feature dimensions (Image 4). The seismic events are arranged spatially into six discrete, tight clusters with little geographic overlap, which represent clearly defined areas of focused activity within the caldera system. Stratification within these clusters is further highlighted by depth analysis, indicating connections with various tectonic and volcanic events. Shallow clusters, which are found between 0 and 10 km below the surface, are probably caused by near-surface deformation connected to caldera structures and volcanic–tectonic faulting. Clusters with an intermediate depth of 10 to 50 km may be indicative of deeper volcanic activities such magma movement, pressurization, or modifications to the subsurface plumbing system. Even though they happen less frequently, deeper occurrences probably represent more extensive regional tectonic impacts, such as forces carried by the Sumatran fault system. DBSCAN successfully captures the horizontal and vertical organization of seismicity, as seen by this depth-dependent clustering, offering insights into the interaction between volcanic and tectonic processes in the Toba Caldera region.

DBSCAN's ability to capture the underlying structure of volcanic seismicity in the Toba region is demonstrated by the clustering visualizations, which show very different spatial organization patterns. Six distinct, extremely small groups, each representing distinct volcanic features, are displayed in the top-panel scatter plot (Figure 3). These clusters verify the algorithm's capacity to

identify coherent seismic zones and are well separated with little spatial overlap, which is consistent with the algorithm's high silhouette score of 0.679. Smaller clusters relate to peripheral volcanic complexes and other structural characteristics, whereas the center cluster, which encompasses the main caldera structure about 98.8° longitude and 2.5° latitude, is the most noticeable. DBSCAN's strengths are further demonstrated by a density overlay (Figure 4,), which shows areas of concentrated activity by precisely identifying the high-density cores of volcanic seismic zones. Simultaneously, the large percentage of points categorized as noise (91.48%) illustrates the pervasive background seismicity characteristic of volcanic–tectonic systems—isolated occurrences or small groupings that do not form significant clusters around key structures. This illustration highlights DBSCAN's capacity to distinguish significant geophysical patterns from dispersed, diffuse activity, which makes it very useful for examining intricate volcanic landscapes.

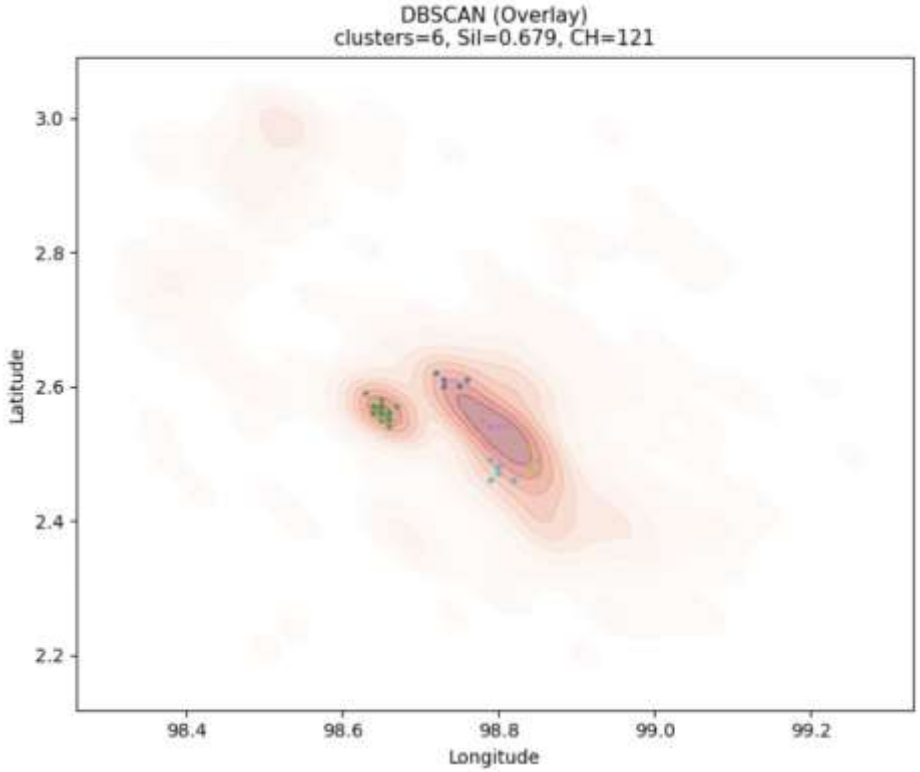

Figure 4. DBSCAN Overlay

With six distinct groups that closely match recognized tectonic and volcanic features, the DBSCAN clustering results show a distinct structure-specific arrangement of seismicity (Figure 5). The core volcanic–tectonic zone associated with caldera dynamics and central volcanic complexes is represented by the orange cluster, which is concentrated around Pangururan and the central caldera rim. The red cluster, which is situated close to Saribu Dolok along the northern caldera rim, most likely represents a unique volcanic complex or a structurally controlled area of the caldera. The brown cluster, which is located in Lake Toba's central-southern region, indicates intra-caldera volcanic activity that may be connected to localized magma migration or resurgent dome processes. Dispersed throughout the caldera's edge, blue clusters are suggestive of secondary volcanic centers or areas impacted by more general regional tectonic stresses. Lastly, the green cluster, which most likely corresponds to caldera rim faulting or other regional structural controls, appears in the eastern sector close to Parapat.

Figure 5. DBSCAN Structure-Specific Identification

Overall, this clustering pattern shows how DBSCAN successfully separates seismic zones of structural and volcanological significance, emphasizing the technique's capacity to directly connect earthquake distributions to geological elements within intricate caldera systems.

CONCLUSIONS

This study emphasizes how clustering algorithms function differently depending on the context in seismological analyses, highlighting the close relationship between algorithm efficacy and the geological features of the study area. With a low Davies-Bouldin index of 0.404 and a high silhouette coefficient of 0.679, DBSCAN works better than other techniques in the Toba Caldera volcanic environment. These measures show that the system can recognize large background activity (which accounts for 91.48% of occurrences) as noise while simultaneously defining six small, structure-specific seismic zones. The discrete, geology-controlled nature of volcanic seismicity where earthquake events are localized around particular caldera structures and volcanic complexes rather than dispersed continuously is what makes DBSCAN effective in this context. This behavior stands in stark contrast to regional tectonic environments, where seismicity frequently follows continuous or linear fault trends and may necessitate the use of different clustering techniques in order to capture structural patterns. These results highlight the significance of taking geological context into account when choosing analytical techniques, showing that algorithm selection should be influenced by both computer efficiency and the underlying physical mechanisms influencing the seismicity.

REFERENCES

[1] C. A. . Chesner, “Toba Caldera : the eye of Sumatra : evolution of the Toba Caldera Complex,” p. 99, 2025.

[2] A. Tibaldi, “Volcano plumbing system geometry: The result of multi-parametric effects,” *J. Volcanol. Geotherm. Res.*, vol. 298, pp. 85–135, Jun. 2015, doi: 10.1016/J.JVOLGEORES.2015.03.023.

[3] J. Saxby, J. Gottsmann, K. Cashman, and E. Gutierrez, “Magma storage in a strike-slip caldera,” *Nat. Commun. 2016 71*, vol. 7, no. 1, pp. 12295-, Jul. 2016, doi: 10.1038/ncomms12295.

[4] A. F. Bell *et al.*, "Caldera resurgence during the 2018 eruption of Sierra Negra volcano, Galápagos Islands," *Nat. Commun.*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/S41467-021-21596-4.

[5] G. Weber, J. Biggs, and C. Annen, "Distinct patterns of volcano deformation for hot and cold magmatic systems," *Nat. Commun.* , vol. 16, no. 1, Dec. 2025, doi: 10.1038/S41467-024-55443-Z.

[6] T. Ryberg, U. Muksin, and K. Bauer, "Ambient seismic noise tomography reveals a hidden caldera and its relation to the Tarutung pull-apart basin at the Sumatran Fault Zone, Indonesia," *J. Volcanol. Geotherm. Res.*, vol. 321, pp. 73–84, Jul. 2016, doi: 10.1016/J.JVOLGEORES.2016.04.035.

[7] M. Sinambela, E. Darnila, I. K. Jaya, I. M. Sarkis, and A. Rikki, "Cluster Analysis and Seismicity of The Samosir Using Machine Learning Approach," *2023 8th Int. Conf. Informatics Comput. ICIC 2023*, 2023, doi: 10.1109/ICIC60109.2023.10381924.

[8] D. H. Natawidjaja, "Updating active fault maps and sliprates along the Sumatran Fault Zone, Indonesia," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 118, no. 1, 2018, doi: 10.1088/1755-1315/118/1/012001.

[9] A. V. H. Simanjuntak *et al.*, "Environmental vulnerability characteristics in an active swarm region," *Glob. J. Environ. Sci. Manag.*, vol. 9, no. 2, pp. 211–226, 2023, doi: 10.22034/gjesm.2023.02.3.

[10] Y. Wang, Y. Gu, and J. Shun, "Theoretically-Efficient and Practical Parallel DBSCAN," *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 2555–2571, Jun. 2020, doi: 10.1145/3318464.3380582/SUPPL_FILE/3318464.3380582.MP4.

[11] Y. P. Wu, J. J. Guo, and X. J. Zhang, "A linear DBSCAN algorithm based on LSH," *Proc. Sixth Int. Conf. Mach. Learn. Cybern. ICMLC 2007*, vol. 5, pp. 2608–2614, 2007, doi: 10.1109/ICMLC.2007.4370588.

[12] D. Z. Chen, M. Smid, and B. Xu, "Geometric algorithms for density-based data clustering," *Int. J. Comput. Geom. Appl.*, vol. 15, no. 3, pp. 239–260, Jun. 2005, doi: 10.1142/S0218195905001683.

[13] P. Viswanath and V. Suresh Babu, "Rough-DBSCAN: A fast hybrid density based clustering method for large data sets," *Pattern Recognit. Lett.*, vol. 30, no. 16, pp. 1477–1488, Dec. 2009, doi: 10.1016/J.PATREC.2009.08.008.

[14] H. Song and J. G. Lee, "RP-DBSCAN: A superfast parallel DBSCAN algorithm based on random partitioning," *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 1173–1187, May 2018, doi: 10.1145/3183713.3196887.

[15] D. Deng, "DBSCAN Clustering Algorithm Based on Density," *Proc. - 2020 7th Int. Forum Electr. Eng. Autom. IFEEA 2020*, pp. 949–953, Sep. 2020, doi: 10.1109/IFEEA51475.2020.00199.

[16] M. Usama *et al.*, "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," *IEEE Access*, vol. 7, pp. 65579–65615, 2019, doi: 10.1109/ACCESS.2019.2916648.

[17] Y. Roh, G. Heo, and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1328–1347, Apr. 2021, doi: 10.1109/TKDE.2019.2946162.

[18] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," *Int. J. Inf. Technol.*, vol. 12, no. 4, pp. 1243–1257, Dec. 2020, doi: 10.1007/S41870-020-00427-7.

[19] W. Jing, C. Zhao, and C. Jiang, "An improvement method of DBSCAN algorithm on cloud computing," *Procedia Comput. Sci.*, vol. 147, pp. 596–604, 2019, doi: 10.1016/J.PROCS.2019.01.208.