Journal Majelis Paspama
Volume 2, Number 2, 2024 (No 108-114)
E.SSN 2987-1379 (Online)

Performance of K-Means Algorithm for Ground Acceleration Clustering

Siska Simamora[1], Amran Manalu[2], Paska Marto Hasugian[3]
[1,2]Fakultas Teknologi dan Bisnis, Universitas Putra Abadi Langkat, [3]Fakultas Ilmu Komputer, Universitas Katolik Santo Thomas

Article Info
Corresponding Author
Siska Simamora
E-mail:
siskamamora22@gmail.com

ABSTRACT
Indonesia is one of the most seismically active regions in the world due to the convergence of the Indo-Australian, Eurasian, and Pacific tectonic plates. This condition exposes the country to frequent earthquakes with varying magnitudes and intensities that may cause severe structural damage and pose risks to human safety. Ground acceleration, particularly Peak Ground Acceleration (PGA), is a key parameter for evaluating earthquake impacts and is strongly influenced by geological conditions, hypocentral depth, and epicentral distance. However, the complexity and large volume of ground acceleration data often hinder manual interpretation. This study applies the K-Means clustering algorithm to classify ground acceleration data obtained from seismic records at several observation points. Prior to clustering, data preprocessing was performed through data cleaning and min–max normalization to ensure quality and comparability across variables. The optimal number of clusters was determined using the Elbow method and Silhouette Score. The results reveal distinct distribution patterns of ground acceleration, which are closely related to local seismic conditions. These findings are expected to contribute to the development of preliminary ground acceleration zonation, providing valuable insights for earthquake hazard mapping and risk mitigation efforts in Indonesia.

Keywords:
Ground Acceleration Clustering, K-Means Algorithm, geological conditions, hypocentral depth, and epicentral distance

INTRODUCTION

Indonesia is one of the countries with the highest seismic activity in the world because it is located at the convergence of three major tectonic plates, namely the Indo-Australian, Eurasian, and Pacific plates. This tectonic setting makes the Indonesian region highly vulnerable to earthquakes that can cause severe damage to infrastructure and threaten human lives. Earthquake events in Indonesia occur frequently and vary in intensity, ranging from small quakes that are barely felt to large, destructive earthquakes. In seismic studies, one of the most important parameters is ground acceleration, which describes the level of shaking intensity at the surface and is highly relevant in assessing potential structural damage. Ground acceleration data are usually obtained from seismic stations distributed across different regions, but these data are highly complex because they are influenced by local geological conditions, distance to the earthquake source, and hypocentral depth. This

complexity makes manual analysis challenging, particularly when dealing with large datasets. Therefore, computational approaches that can systematically organize and classify the data are needed to better understand the distribution patterns of ground acceleration.

The complexity of ground acceleration data highlights the importance of analytical methods that can simplify and reveal hidden patterns in large datasets. One promising approach is clustering analysis, which groups data into clusters based on similarity in their characteristics. By applying clustering, data that initially appear random can be structured into more meaningful groups, allowing for easier interpretation and analysis. In the context of seismic hazards, clustering can be used to identify regions with similar ground acceleration patterns, which may be associated with shared geological characteristics or comparable distances from earthquake sources. This approach is particularly valuable for seismic hazard zoning because regions belonging to the same cluster can be managed with similar mitigation strategies. Consequently, clustering not only facilitates scientific interpretation but also provides practical insights, especially for disaster mitigation planning that requires data-driven decision-making.

Among the various clustering methods, the K-Means algorithm is one of the most widely used and effective approaches. K-Means belongs to the category of unsupervised learning because it does not rely on pre-labeled data or predefined targets. The algorithm works by partitioning the dataset into a predefined number of clusters, beginning with the selection of initial centroids for each cluster. Each data point is then assigned to the nearest centroid, and the centroid positions are recalculated as the mean of the assigned data points. This iterative process continues until the cluster assignments stabilize. The advantages of K-Means include its simplicity, computational efficiency, and effectiveness in handling large, high-dimensional datasets. In geophysics, K-Means has been successfully applied to seismic and geophysical data to identify patterns that are difficult to detect through manual observation. Several previous studies have demonstrated that the algorithm is capable of uncovering meaningful structures in seismic data, thereby providing valuable insights for understanding earthquake phenomena and associated hazards.

Based on this background, the present study focuses on the application of the K-Means clustering algorithm to classify ground acceleration data in Indonesia. The main objective is to identify clearer and more systematic distribution patterns that can enhance the understanding of seismic characteristics in different regions. The expected outcome of this research is to provide a solid foundation for the development of more accurate seismic hazard zoning maps. These maps can serve as references for earthquake-resistant infrastructure development, land-use planning, and disaster mitigation strategies. Thus, the application of K-Means clustering not only contributes academically to data analysis in seismology but also offers practical benefits in protecting society from the risks of earthquakes. Ultimately, integrating geoscience with modern computational approaches such as K-Means clustering has the potential to open new opportunities for seismic risk management in Indonesia, a country that remains highly vulnerable to seismic hazards.

.

METODE

Research Data

The data used in this study consist of ground acceleration values obtained from seismic records at several observation points. Each record contains the Peak Ground Acceleration (PGA) along with supporting parameters such as earthquake magnitude and epicentral

distance. Prior to analysis, the dataset was cleaned by removing extreme or irrelevant values (data cleaning) to ensure the reliability and quality of the data.

Data Preprocessing

Before clustering, the ground acceleration data were normalized using the min–max normalization method so that each variable falls within the range of 0–1. This step is essential to prevent differences in variable scales from disproportionately influencing the distance calculations in the K-Means algorithm.

K-Means Algorithm

The K-Means method is a partition-based clustering technique that divides data into *K* groups by minimizing the distance between data points and their corresponding cluster centers (centroids). The steps of the K-Means algorithm applied in this study are as follows:

1. Determining the number of clusters (*K*) to be used.
2. Randomly initializing the initial centroids.
3. Calculating the distance of each data point to the centroids using Euclidean distance.
4. Assigning each data point to the cluster with the nearest centroid.
5. Updating the centroid positions based on the mean of the data points within each cluster.
6. Repeating steps 3–5 until the centroids stabilize or the maximum number of iterations is reached.

Determining the Optimal Number of Clusters

The optimal number of clusters was determined using the Elbow method and the Silhouette Score. The Elbow method identifies the point of significant change in the inertia curve (the sum of squared distances within clusters), while the Silhouette Score evaluates how well the clusters are separated from one another.

Clustering Evaluation

The results of the ground acceleration clustering were analyzed to examine the data distribution within each cluster. Further interpretation was carried out by associating the clustering outcomes with the seismic conditions of the observation areas. These results are expected to provide an initial overview of ground acceleration zonation, which can serve as valuable input for earthquake risk mitigation strategies.

RESULTS AND DISCUSSION

Clustering using the K-Means algorithm essentially groups data into clusters based on their proximity or similarity. This process begins by randomly determining a number of initial centroids, which will serve as the center of each cluster. Each data item in the dataset is then calculated for its distance from these centroids, and the data is placed in the cluster with the closest centroid.

This process is carried out iteratively, where each iteration updates the centroid position based on the average position of the data within its cluster. Thus, the centroid will shift position to more accurately represent the center of the data within that cluster. This iteration continues until there are no more significant changes in the centroid position, indicating that the clustering process has reached convergence. In testing using K-Means, one of the main parameters is the number of clusters to be formed. This number of clusters must be determined in advance before the clustering process begins. Additionally, K-Means also uses a distance metric, usually the Euclidean distance, to calculate the closeness between the data and the centroid.

In this test, the dataset used is 1457 data, and it will be divided into 5 clusters. The initial process uses randomly selected centroids, and the iterative process is carried out until convergence is achieved, namely when there are no more significant changes in the division of data into clusters. The results of the execution of clustering formation with K-Means with FormationCenter point for the first data randomly and the next point using the average function, with the following center point information:

Iteration	Center 1 (Depth, Mag, Dmin, PGA)	Center 2 (Depth, Mag, Dmin, PGA)	Center 3 (Depth, Mag, Dmin, PGA)	Center 4 (Depth, Mag, Dmin, PGA)	Center 5 (Depth, Mag, Dmin, PGA)
Beginning	35, 5.3, 2011, 642.48	23613, 4.4, 3122, 15.92	10, 4.7, 2324, 394.16	145458, 4.4, 2917, 1.46	60.52, 4.6, 1288, 68.59
Iteration 1	1804.55, 4.55, 1900, 428.38	45851.64, 4.53, 1638, 10.28	686.41, 4.51, 3241, 267.44	192936, 4.39, 1795, 1.44	1716.28, 4.55, 868, 56500.27
Iteration 2	7925.03, 4.57, 1659, 480.86	62578.04, 4.49, 1659, 6.22	35.70, 4.53, 2782, 355.80	224140, 4.38, 1769, 1.10	22.91, 4.53, 1.65, 79699.93
Iteration 3	18712.01, 4.52, 1829, 32.51	79044.84, 4.48, 1646, 4.52	35.74, 4.54, 2390, 1197.80	256661, 4.37, 1834, 0.89	12.61, 4.55, 1.69, 90877.92
Iteration 4	27918.08, 4.54, 1822, 21.66	98590.67, 4.44, 1639, 3.19	373.61, 4.54, 2355, 1310.49	296000, 4.37, 1836, 0.72	11.88, 4.55, 1.68, 92497.50
Iteration 5	40912.49, 4.55, 1724, 10.68	121239.68, 4.41, 1632, 2.21	1647.12, 4.53, 2298, 1154.66	337863, 4.37, 1991, 0.58	11.88, 4.55, 1.68, 92497.50
Iteration 6	49081.63, 4.52, 1600, 7.95	140930.24, 4.39, 1701, 1.72	2588.96, 4.54, 2290, 1089.85	382184, 4.40, 2127, 0.49	11.88, 4.55, 1.68, 92497.50
Iteration 7	55047.99, 4.51, 1623, 6.67	154158.74, 4.38, 1692, 1.50	3652.36, 4.54, 2255, 1036.24	430292, 4.42, 2251, 0.40	11.88, 4.55, 1.68, 92497.50
Iteration 61	115564.63, 4.40, 1701, 2.24	222729.46, 4.34, 1652, 0.86	17313.18, 4.54, 2063, 711.18	507619, 4.47, 2303, 0.30	11.88, 4.55, 1.68, 92497.50

After the center point is formed, the calculation process is carried out for each iteration in this experiment with a maximum of 100 iterations.This process continues until there is no significant change in the data distribution, or the maximum number of iterations is reached. Euclidean distance is used to measure closeness at each iteration. The following describes the clustering process using K-Means, where each data point is tested in each iteration to determine the most appropriate cluster. The clustering equation for each iteration is described below:

Table 1 K-Means Clustering Iterations 1-7

No	Depth	Mag	Dmin	PGA	I-1	I-2	I-3	I-4	I-5	I-6	I-7
1	10187	5.6	2366	85.82	1	1	1	1	3	3	3
2	11326	4.6	2683	45.21	1	1	1	1	3	3	3
3	29.51	4.5	2634	302.79	3	3	3	3	3	3	3
4	9221	4.8	3743	61.41	3	1	1	3	3	3	3
5	54874	4.9	1945	6.79	2	2	2	2	1	1	1
6	204591	4.2	1202	0.84	4	4	4	4	4	2	2
7	233212	5	5.46	1,058	4	4	4	4	4	4	2
8	63029	4.7	4546	5,106	2	2	2	2	1	1	1
9	46177	4.6	0.971	7,348	2	2	2	1	1	1	1
10	134961	4.8	1857	1,972	4	4	2	2	2	2	2
...	...	...	...	...	...	...	...	...	...	...	...
1538	258863	4	2.31	0.559	4	4	4	4	4	4	2
1539	117.6	5.2	1873	668.88	1	3	3	3	3	3	3
1540	31319	4	1288	9,074	2	2	1	1	1	1	1
1541	21087	5.1	2427	26.29	2	1	1	1	1	3	3
1542	184158	4.2	0.763	0.969	4	4	4	4	2	2	2
1543	42226	4.1	1294	6,435	2	2	2	1	1	1	1
1544	145614	4.3	1914	1,389	4	4	4	2	2	2	2
1545	35	4.4	2743	273.14	3	3	3	3	3	3	3
1546	146223	4.3	1749	1,3817	4	4	4	2	2	2	2
1547	10065	4.2	3634	41,351	3	1	1	1	3	3	3

Visualization of data clustering with K-Means is presented in the following table and graph.This :

Table 2 Number of Data for Each Cluster

Iteration/ Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
I-1	123	450	278	458	238
I-2	255	501	375	341	75
I-3	219	519	495	252	62
I-4	309	478	518	182	60
I-5	356	407	591	133	60
I-6	399	363	627	98	60
I-7	406	348	660	73	60

Data clustering using the K-Means algorithm is shown in the following graph, which shows the distribution of data into several clusters according to the proximity of their characteristics.

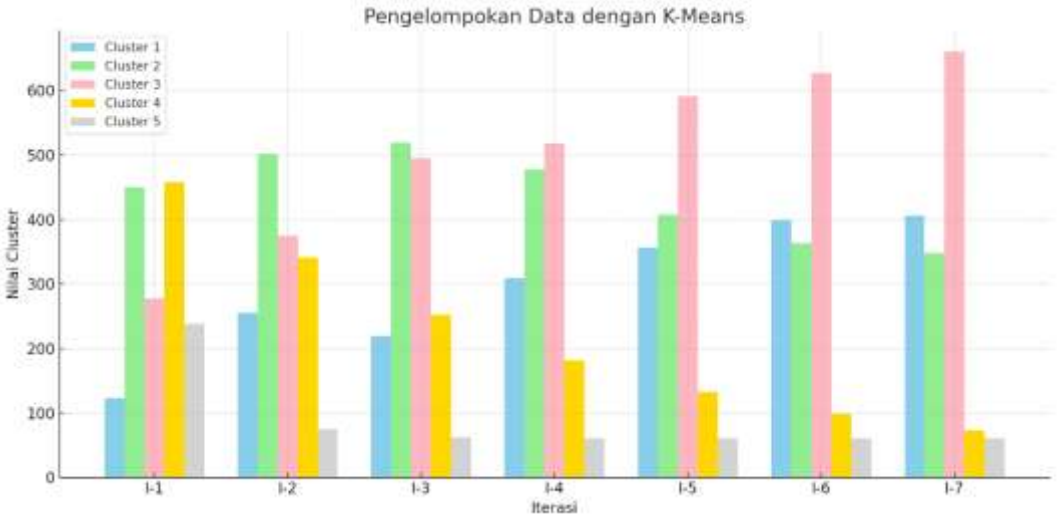

Figure 1 Visualization of K-Means Clustering for 7 Iterations

Picture1. shows the results of data clustering using the K-Means Algorithm, which divides the data into five clusters based on iterations from I-1 to I-7. Each bar represents the number of data included in each cluster at each iteration. Cluster 1 experienced an increase from 123 to 406 data at the end of the iteration, while Cluster 3 grew significantly, reaching 660 data in the final iteration. In contrast, Cluster 4 and Cluster 5 showed a decrease in the number of data, with Cluster 5 remaining stable at 60 data from iterations I-4 to I-7. This visualization shows the dynamics of changes in the number of data in each cluster during the iteration process.

Discussion

The clustering process using the K-Means algorithm on the earthquake dataset was carried out by determining the number of clusters to five, where the initial centroid was randomly selected to begin the clustering process. Each data item was calculated for its distance from the centroid using the Euclidean distance, then placed in the cluster with the closest center. After that, the centroid was updated based on the average position of the data within the cluster, and this step was repeated iteratively until convergence was achieved. The experimental results showed that in the initial iterations the number of data between clusters was not balanced, for example, Cluster 1 only contained 123 data items while Cluster 2 contained 450 data items. However, as the iterations increased, the data distribution began to find a more stable pattern. Cluster 1 experienced a consistent increase to 406 data items, while Cluster 3 grew significantly to 660 data items at the end of the iteration. Conversely, Cluster 4 experienced a decrease in the number of members from 458 to 73 data items, while Cluster 5 reached stability more quickly and remained containing 60 data items since the fourth iteration.

The change in the amount of data in each cluster indicates that the K-Means process gradually moves the data into the most appropriate groups until the centroid position truly represents the center of the distribution. Large clusters are formed in groups with common data characteristics, while small clusters serve to accommodate data with extreme values or special characteristics, such as very low depths or high ground acceleration. After 61 iterations, the centroid position remains relatively stable, with very little change in the data distribution, thus converging the algorithm before the maximum iteration limit is reached.

Overall, the clustering results confirm that K-Means is able to divide the earthquake data into five distinct groups with a tendency towards uneven distribution. Two dominant clusters contain the majority of the data, while the other clusters are smaller and serve to group specific data. This condition demonstrates the algorithm's effectiveness in revealing the structure of the data distribution while providing important information about homogeneous and anomalous data groups.

CONCLUSION

The application of the K-Means algorithm to an earthquake dataset with 1,457 data sets and five clusters demonstrates that the iterative process is capable of producing stable and convergent clustering. The initially randomly selected centroid gradually moves until it more accurately represents the center of the data distribution. Changes in the number of members between clusters during iterations demonstrate the dynamics of data redistribution, where Cluster 1 and Cluster 3 become the dominant groups with the largest data sets, while Cluster 4 and Cluster 5 experience a decrease and stabilize at a small number. This condition confirms that the K-Means algorithm is effective in revealing data distribution patterns, separating groups with common characteristics from those with extreme or anomalous characteristics. By achieving convergence before the maximum iteration limit, the clustering results provide a clear picture of the earthquake data structure, both for large, homogeneous groups and small groups reflecting specific characteristics.

REFERENCE

Sonkar, R., Dhekne, P. Y., & Londhe, N. D. (2021). Improvement in the prediction of peak particle velocity of blast-induced ground vibrations using K-means clustering. *Arabian Journal of Geosciences*, *14*(22), 2255.

Li, Y., Zhou, X., Gu, J., Guo, K., & Deng, W. (2022). A novel K-means clustering method for locating urban hotspots based on hybrid heuristic initialization. *Applied Sciences*, *12*(16), 8047.

Pangestu, M. S., & Fitriani, M. A. (2022). Perbandingan Perhitungan Jarak Euclidean Distance, Manhattan Distance, dan Cosine Similarity dalam Pengelompokan Data Bibit Padi Menggunakan Algoritma K-Means. *Sainteks*, *19*(2), 141-155.

Hartono, B., Eniyati, S., & Hadiono, K. (2023). Perbandingan Metode Perhitungan Jarak pada Nilai Centroid dan Pengelompokan Data Menggunakan K-Means Clustering. *Jurnal Sistem Komputer dan Informatika (JSON) Hal*, *503*, 509.

Prasetiani, S. D., & Rochmawati, N. (2022). Penerapan Data Mining Untuk Clustering Menu Favorit Menggunakan Algoritma K-Means (Studi Kasus Kedai Expo). *Journal of Informatics and Computer Science (JINACS)*, *3*(03), 278-286.

Kartikawati, L. (2022). Analisis Kualitas Pengelompokkan Algoritma K-Means di Knime dan Excel untuk PTMT Pasca Vaksinasi Covid-19. *Ideguru: Jurnal Karya Ilmiah Guru*, *7*(1), 70-79.

Widyatami, A. I., & Reistiani, V. M. (2023, October). Clustering Wilayah Potensi dan Strategi Pengembangan Komoditas Unggulan Tanaman Hortikultura dan Palawija Level Kecamatan di Sumatera Barat Tahun 2021. In *Seminar Nasional Official Statistics* (Vol. 2023, No. 1, pp. 41-52).

Zet, L., Yuniarti, E., Azzahra, S. F., Laia, T. K., & Wulandari, R. (2023). *Monograf Akselerasi Pembangunan Beserta Dampak Lingkungannya*. Mega Press Nusantara.