International Student Research Review
Volume 2 Issue 1 May 2025
DOI: https://doi.org/10.53378/isrr.163

Enhancement of K-Means algorithm for analyzing earthquake occurrence pattern in the Philippines

[1]Sean Marie Bayono, [2]Ronanne Jcher Bulaon, [3]Richard C. Regala, [4]Vivien A. Agustin & [5]Khatalyn E. Mata

Abstract

This study aims to enhance the K-Means clustering algorithm to improve the analysis of earthquake occurrence patterns in the Philippines. Traditional K-Means, while effective, suffers from limitations such as random initialization and slow convergence. To address these issues, we propose an improved K-Means algorithm that strategically selects initial centroids based on a distance-weighted probability distribution to enhance accuracy and processes data in smaller batches to reduce computation time, thereby improving scalability and convergence speed. Using earthquake data from the Philippine Institute of Volcanology and Seismology (PHIVOLCS), we evaluate the performance of the enhanced algorithm using metrics such as Silhouette Score and Time Complexity. Results demonstrate that the proposed modifications significantly enhance clustering accuracy, computational efficiency, and scalability, leading to more precise identification of high-risk seismic areas. By providing a more accurate and efficient framework for seismic data analysis, this research contributes to disaster preparedness, risk mitigation, and informed decision-making in urban planning and disaster management.

Keywords: *K-Means algorithm, mini-batch processing, disaster preparedness, seismic data analysis*

Article History:

***Received**: February 17, 2025* ***Revised**: March 8, 2025*
***Accepted**: March 11, 2025* ***Published online**: May 31, 2025*

Suggested Citation:
Bayono, S.M., Bulaon, R.J., Regala, R.C., Agustin, V.A. & Mata, K.E. (2025). Enhancement of K-Means algorithm for analyzing earthquake occurrence pattern in the Philippines. *International Student Research Review*, 2(1), 1-18. https://doi.org/10.53378/isrr.163

About the authors:
[1]Corresponding author. Undergraduate student. Department of Computer Science. College of Information Systems and Technology Management - Pamantasan ng Lungsod ng Maynila. Email: smbbayono2021@plm.edu.ph
[2]Undergraduate student. Department of Computer Science. College of Information Systems and Technology Management - Pamantasan ng Lungsod ng Maynila
[3]Bachelor's Degree in Information Communication Technology. Pamantasan ng Lungsod ng Maynila. Computer Laboratory Administrator
[4]Master in Information Technology. College of Information Systems and Technology Management - Pamantasan ng Lungsod ng Maynila. Associate Dean/Assistant Professor III
[5]Doctor in Information Technology. Dean - College of Information Systems and Technology Management, Pamantasan ng Lungsod ng Maynila.

1. Introduction

Earthquakes are natural phenomena that can cause significant destruction and loss of life, making it crucial to analyze their occurrence patterns for better preparedness and risk mitigation. Understanding the spatial and temporal distribution of earthquakes helps scientists, policymakers, and disaster response teams to identify high-risk areas, develop effective emergency response plans, and implement safety measures to minimize the impact of seismic events. By analyzing earthquake patterns, data-driven decisions can be made to enhance public safety and improve institutional effectiveness in disaster management.

Data clustering is a powerful method for analyzing earthquake occurrences, as it groups events with similar characteristics, providing deeper insights into seismic behaviors and trends (Fan & Xu, 2019). K-Means algorithm is widely used for grouping data among the clustering algorithms due to its simplicity and efficiency (Xu et al., 2013). However, despite its effectiveness, K-Means has limitations such as sensitivity to initial centroid placement, convergence speed, and computational complexity when handling large datasets. To address these limitations, this study aims to enhance the K-Means algorithm by introducing improvements that optimize its computational efficiency and initialization process. These enhancements involve processing data in smaller subsets to reduce computation time while maintaining clustering accuracy and employing advanced methods for initializing cluster centroids to improve the algorithm's accuracy and speed of convergence (Béjar, 2020).

By refining the algorithm, the study aims to improve the K-Means algorithm to address its existing challenges and apply it to the analysis of earthquake occurrence patterns in the Philippines. By refining the algorithm, this research aims to provide more accurate and efficient clustering of earthquake data, utilizing the open-source data provided by the Philippine Institute of Volcanology and Seismology (PHIVOLCS). The enhanced clustering approach will help identify high-risk areas, tailor emergency interventions, and improve disaster preparedness and response strategies.

This research contributes significantly to seismic data analysis by refining an existing clustering algorithm for practical application in earthquake studies. The enhanced K-Means algorithm will serve as a more reliable tool for analyzing seismic data, allowing scientists and policymakers to make better-informed decisions. Additionally, the findings of this study

will advance the field of seismic data mining and clustering, laying the groundwork for future research and applications in disaster risk management.

2. Literature review

Clustering algorithms are fundamental tools in data analysis, widely used to identify patterns and group similar data points based on their characteristics. These algorithms are applied across various fields, including market research, image segmentation, and seismic data analysis, making them indispensable for extracting meaningful insights from complex datasets. Clustering techniques can generally be categorized into partitioning, hierarchical, density-based, and grid-based methods, each offering unique advantages depending on the nature of the data and analytical goals (Xu & Tian, 2015). Among these, partitioning methods like the K-Means algorithm are particularly popular due to their simplicity and efficiency. However, traditional clustering methods often face challenges such as sensitivity to noise, difficulty in handling irregular cluster shapes, and determining the optimal number of clusters (Rousseeuw, 1987). These limitations have prompted researchers to explore advanced variations and hybrid algorithms that can adapt to the specific characteristics of the data, ensuring more robust and accurate results.

In seismic data analysis, clustering is critical in understanding earthquake occurrence patterns, which is essential for disaster preparedness and risk mitigation. The K-Means algorithm has been extensively used in seismic data analysis due to its ability to group earthquake occurrences based on attributes such as magnitude, depth, and geographic location. For instance, Fan and Xu (2019) applied K-Means and DBSCAN to analyze seismic data, revealing earthquake distribution patterns and improving disaster preparedness. The study demonstrates that while K-Means are effective for convex-shaped clusters, they struggle with the irregular clustering of seismic belts. In contrast, DBSCAN, which accommodates clusters of arbitrary shapes, excels in fitting seismic belts, making it a valuable tool for seismic distribution analysis. Similarly, Novianti et al. (2017) employed K-Means to analyze earthquake epicenters in Bengkulu Province, Indonesia, identifying seven distinct seismic zones using the Krzanowski and Lai (KL) index. Their study highlighted that while K-Means successfully classified seismic zones, they struggled with irregular cluster shapes and noise. This limitation emphasizes the need for enhancements that improve clustering accuracy and robustness.

Despite its widespread use, K-Means faces several challenges that limit its effectiveness in complex datasets. One major issue with traditional K-Means is its sensitivity to initial centroid placement, which can lead to suboptimal clustering results and convergence to local minima (Celebi et al., 2013). Mato and Toulkeridis (2017) introduced the GEO K-Means algorithm to address the anisotropic nature of geophysical data, achieving a more precise classification of seismic events by incorporating contextual information about seismic precursors. Similarly, Rifa et al. (2020) compared K-Means with K-Medoids for clustering earthquake data in Indonesia, finding that while K-Means efficiently classifies data, it is susceptible to outliers and noise. K-Medoids, which uses medoids instead of centroids, offers a more robust alternative, suggesting that hybrid approaches could enhance K-Means' reliability.

Another significant challenge of K-Means is its computational inefficiency when handling large datasets. Traditional K-Means require repeated reassignments of data points to centroids, leading to high computational costs. Sculley (2010) addressed this by introducing Mini-Batch K-Means, which processes small, randomly selected data samples at each iteration, significantly reducing computational overhead and improving convergence rates. While Mini-Batch K-Means enhance scalability, Béjar (2020) and Hicks et al. (2021) observed that it may compromise clustering accuracy due to its stochastic nature. To address this, Xiao et al. (2018) proposed the SMK-Means algorithm, which integrates simulated annealing optimization with Mini-Batch K-Means to improve clustering precision, particularly in anomaly detection scenarios. This approach prevents the algorithm from getting trapped in local optima, illustrating the potential of combining optimization techniques to enhance traditional K-Means.

Despite these advancements, a critical research gap remains. Many enhancements, such as K-Means++ for centroid initialization and Mini-Batch K-Means for improved computational efficiency, address specific limitations but do not fully resolve the challenges of balancing clustering accuracy, robustness, and efficiency in large-scale seismic data analysis. The random initialization of centroids in K-Means often leads to inconsistent results, while Mini-Batch K-Means sacrifices accuracy for speed. Furthermore, existing clustering methods struggle with seismic data's dynamic and irregular nature, which often contains noise and outliers and does not conform to spherical cluster shapes.

This study aims to bridge these gaps by proposing an Enhanced K-Means algorithm that integrates two key improvements: distance-weighted centroid initialization and mini-batch processing. The distance-weighted centroid initialization method ensures better centroid placement, reducing sensitivity to initial configurations and improving clustering consistency. The mini-batch processing approach enhances computational efficiency by reducing the data processed per iteration while maintaining clustering quality. These modifications collectively aim to improve clustering accuracy and scalability, particularly for analyzing earthquake occurrence patterns in large seismic datasets.

By addressing the limitations of traditional and existing enhanced K-Means algorithms, this research contributes to seismic data clustering by offering a more reliable, accurate, and computationally efficient method. The proposed modifications provide a more robust framework for identifying high-risk seismic zones, enhancing disaster preparedness, and supporting informed decision-making in urban planning and disaster risk reduction efforts.

3. Methodology

3.1. Existing K-Means Algorithm

1: uniformly sample k points $c_1, \ldots, c_k$ from X, let $C = \{ c_1, \ldots, c_k \}$

2: repeat

3: for each $i \in \{1, \ldots, k\}$ do

4: $$C_i = \{x_j | c_i = \arg min_{c_i \in C} d(x_j, C)\}$$

5: end for

6: for each $i \in \{1, \ldots, k\}$ do

7: $$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

8: end for

9: until convergence

The K-Means algorithm is a widely used clustering technique for partitioning a dataset into k clusters. It begins by selecting k initial centroids, typically chosen at random. Each data point is then assigned to the cluster whose centroid is the nearest, as measured by a distance metric, typically the Euclidean distance. Once all points are assigned, the centroids are updated by computing the mean of all points within each cluster. This process of

assignment and update is repeated iteratively until the centroids no longer change significantly, indicating convergence. K-Means is efficient and easy to implement but can be sensitive to the initial centroid placement and may converge to local optima.

3.2. Equations for Enhanced Algorithm

The study used the following equations to enhance the K-Means algorithm. The K-Means++ seeding technique is applied to improve the standard random initialization of centroids. Poor random initialization often leads to suboptimal clustering results, affecting the robustness and convergence of the K-Means algorithm (Arthur & Vassilvitskii, 2007). To address this issue, K-Means++ aims to spread out the initial centroids more effectively by selecting them based on a distance-weighted probability distribution (1).

$$P(x) = \frac{D(x)^2}{\sum_{x' \in X} D(x')^2} \quad (1)$$

Where:

x = data point in the dataset X

$D(x)^2$ = the squared Euclidean distance between x and its nearest previously chosen centroid

C = the set of centroids selected so far

$\sum_{x' \in X} D(x')^2$ = sum of the squared distances for all points in X

The Euclidean distance is a fundamental distance metric in clustering algorithms like K-Means. It measures the straight-line distance between two points in a multidimensional space, providing a simple yet effective way to quantify similarity between data points (Han et al., 2011). In the K-Means++ initialization, this metric is crucial in determining the probability distribution for selecting initial centroids. Specifically, the distance-weighted probability distribution formula (1) uses the squared Euclidean distance to ensure that the algorithm selects new centroids that are more likely to be farther from existing centroids.

The squared Euclidean distance, given in equation (2), is defined as:

$$D(x) = \min_{c \in C} ||x - c||^2 \quad (2)$$

Where:

$||x - c||$ = Euclidean distance between point x and centroid c

$\min_{c \in C}$ = the minimum distance to the closest selected centroid in set C

To improve the runtime of the standard K-Means algorithm, the study integrated the mini-batch centroid updating method, which processes smaller subsets of data at each iteration rather than the entire dataset. This approach significantly reduces computational complexity by limiting the number of distance calculations and updates performed per iteration, making it particularly effective for large-scale datasets (Sculley, 2010). Unlike the traditional K-Means algorithm, which requires recalculating centroids based on all data points in each iteration, mini-batch K-Means updates centroids incrementally, leading to faster convergence while maintaining comparable clustering quality (Bottou & Bousquet, 2008). Additionally, this enhancement incorporates a learning rate η (3) and gradient step (4), allowing centroids to adjust dynamically based on the mini-batch samples. This technique prevents drastic centroid shifts and helps stabilize the optimization process, ultimately improving efficiency and scalability (Dean et al., 2012).

The learning rate η (3), adapts based on the number of times a centroid has been updated.

$$\eta = \frac{1}{v[c]} \tag{3}$$

Where:

η = learning rate

$v[c]$ = number of times the centroid c has been updated

The gradient step (4) updates the centroids based on the mini-batch data.

$$C = (1 - \eta)c + \eta x \tag{4}$$

Where:

C = updated position of centroid

c = current position of centroid

η = learning rate

x = a data point from the current mini-batch

3.3. Evaluation Metrics Used

The study used the silhouette score as an evaluation metric to comprehensively measure clustering quality by balancing cohesion and separation. It measures how similar an object is to its cluster (cohesion) compared to other clusters (separation) (Shahapure & Nicholas, 2020). The score helps determine whether data points have been clustered correctly and can indicate whether a clustering algorithm effectively groups similar data points together while separating different clusters. Analyzing how well each data point fits within its assigned cluster relative to others, the Silhouette Score ensures that the algorithm correctly groups similar data while maintaining clear cluster boundaries.

$$s = \frac{(b - a)}{\max(a, b)} \tag{5}$$

Where:

s = silhouette score

a = average intra-cluster distance (cohesion measure)

b = average nearest-cluster distance (separation measure)

This formula results in scores that typically range between -1 and +1. A value near +1 indicates that the sample is well-clustered. A value near 0 means the sample is on or very close to the decision boundary between clusters. A value near -1 implies that the sample may be in the wrong cluster (Shahapure & Nicholas, 2020).

The study used the time complexity metric to evaluate the computational efficiency of centroid updates in the clustering process. This metric is essential for evaluating the runtime efficiency of centroid updating in clustering algorithms. It calculates the computational cost by considering the number of clusters, data points per update, and iteration count. This metric assesses whether the algorithm's runtime has been optimized (Xu & Tian, 2015).

$$T = k \times n \times t \tag{6}$$

Where:

T = time complexity

k = number of clusters

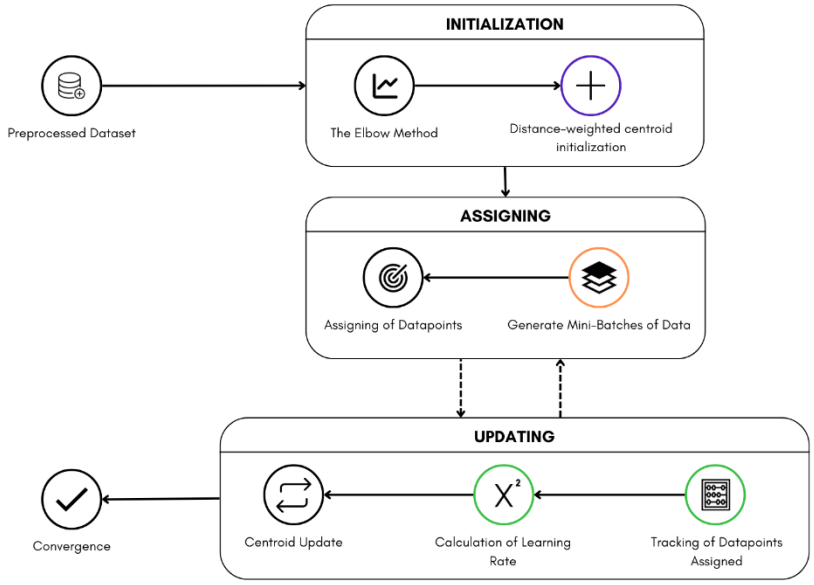
n = number of data points used per update

t = number of iterations for centroid update

3.4. Enhanced Algorithm Conceptual Framework

Figure 1

Conceptual Framework of Enhanced K-Means Algorithm



The conceptual framework for the Enhanced K-Means Algorithm, illustrated in figure 1, outlines the modifications made to improve the efficiency, scalability, and accuracy of traditional K-Means clustering. This framework consists of three fundamental phases: initialization, assigning, and updating, each addressing specific challenges associated with the original algorithm. The algorithm begins with a preprocessed dataset to remove irrelevant attributes, address missing or null values, and retain only the attributes from the dataset.

In the initialization phase, the optimal number of clusters is determined using the elbow method. Then, it integrates a distance-weighted centroid initialization technique to enhance the centroid selection process. This method, inspired by K-Means++, reduces the algorithm's sensitivity to initial centroid placement by ensuring a more effective distribution of centroids across the dataset. It then applies a distance-weighted centroid initialization technique to distribute centroids more effectively, reducing sensitivity to initial placement and improving convergence stability. During the assigning phase, the algorithm allocates data points to clusters using mini-batch processing, which processes smaller subsets of data iteratively instead of analyzing the entire dataset at once. Mini-batch processing significantly reduces computational overhead and memory usage, enhancing the algorithm's efficiency for

large datasets while maintaining clustering accuracy. In the updating phase, the algorithm recalculates centroid positions incrementally using a learning rate and a gradient step, optimizing the rate at which centroids move toward their optimal positions. The tracking mechanism monitors data point assignments, ensuring stability in cluster allocations throughout iterations. The algorithm continues this process until convergence, where further updates no longer produce significant changes.

3.5. Tools

Google Colab is used as the primary development environment used for the enhancement of the K-Means Algorithm. Google Colab is a cloud-based platform that enables researchers to write and execute Python code in a browser with the added benefit of access to high-performance computing resources, such as GPUs and TPUs, for free. It offers built-in support for popular libraries and frameworks, making it an ideal environment for machine learning, data analysis, and various Python-based computations.

Python was chosen as the programming language due to its extensive ecosystem of scientific libraries, which were crucial for implementing and optimizing the K-Means algorithm in this study. The research focused on improving the traditional K-Means algorithm by incorporating enhancements such as a distance-weighted initialization process for selecting initial centroids and mini-batch processing for updating centroids incrementally based on smaller data subsets. These modifications were implemented to enhance clustering performance and computational efficiency. The following libraries were used for enhancing the algorithm and evaluating its performance:

kneed 0.7.0 – used to identify the optimal number of clusters using the elbow method.

Matplotlib 3.6.0 – for visualizing data, including cluster results and the elbow plot.

Scikit-learn 1.2.0 – a comprehensive library for machine learning tasks, used for clustering, dataset generation, and model evaluation.

Kneed's KneeLocator – for detecting the knee/elbow point in data curves.

StandardScaler from Scikit-learn – used for standardizing the features before clustering.

3.6. Dataset and Data Processing

The dataset used in this study is derived from the open-source records of the PHIVOLCS, specifically focusing on earthquake occurrences in the Philippines. The data

spans various time frames to evaluate the scalability and performance of the analysis. It is divided into four separate files: the first file contains three months of data, from July 2024 to September 2024; the second file covers six months, from April 2024 to September 2024; the third file spans nine months, from January 2024 to September 2024; and the final file includes twelve months of data, from October 2023 to September 2024. Each dataset highlights the following attributes: date-time, longitude (°E), latitude (°N), depth (km), magnitude, and location.

To prepare the data for analysis, the datasets are converted into comma-separated values (CSV) files and undergo preprocessing, which includes cleaning, normalizing, and transforming the data. This involves removing irrelevant attributes, addressing missing or null values, and retaining only the attributes relevant to the analysis of earthquake occurrences. By utilizing datasets of varying volumes, the study aims to assess the scalability and overall performance of the algorithm across different data sizes.

Enhanced Algorithm

Enhanced K-Means Algorithm

Initialization

Apply Distance-Weighted Probability Distribution

Select randomly the first centroid from the data points.

For each remaining centroid, choose a new data point with a probability proportional to its distance squared from the closest already chosen centroid.

Assigning

Evenly distribute the datapoints to mini-batches based on the entire dataset.

For each data point in the mini-batch, find the nearest cluster centroid for the datapoint.

Store the datapoint to its nearest cluster centroid.

Updating

For each datapoint in the mini batch, retrieve the stored datapoints.

Track the count of data points assigned to each centroid.

Calculate the learning rate.

Update centroid using based on the learning rate and the datapoint.

Repeat steps 2 and 3 until convergence.

Analysis and Visualization

After clustering the earthquake occurrences dataset, the clustered data were visualized and analyzed to extract valuable information, including the intensity of earthquakes in specific areas and the frequency of their occurrences.

4. Findings and Discussion

4.1. Dataset Preprocessing Results

The dataset acquired from the open-source records of the PHIVOLCS spanning a 12-month period was analyzed and preprocessed to address missing values and ensure data quality. The dataset exhibited substantial missing values such as date - time (20,969 missing values), location (20,969 missing values), and unnamed columns (Unnamed: 6, Unnamed: 7, and Unnamed: 8), likely due to non-numeric or empty entries. These columns were dropped as their missing values exceeded the 50% threshold. Minor missing values were also identified in depth (km) (2 missing values) and were filled with the column mean. The remaining columns, including latitude (ºN), longitude (ºE), and magnitude, were fully populated with no missing values.

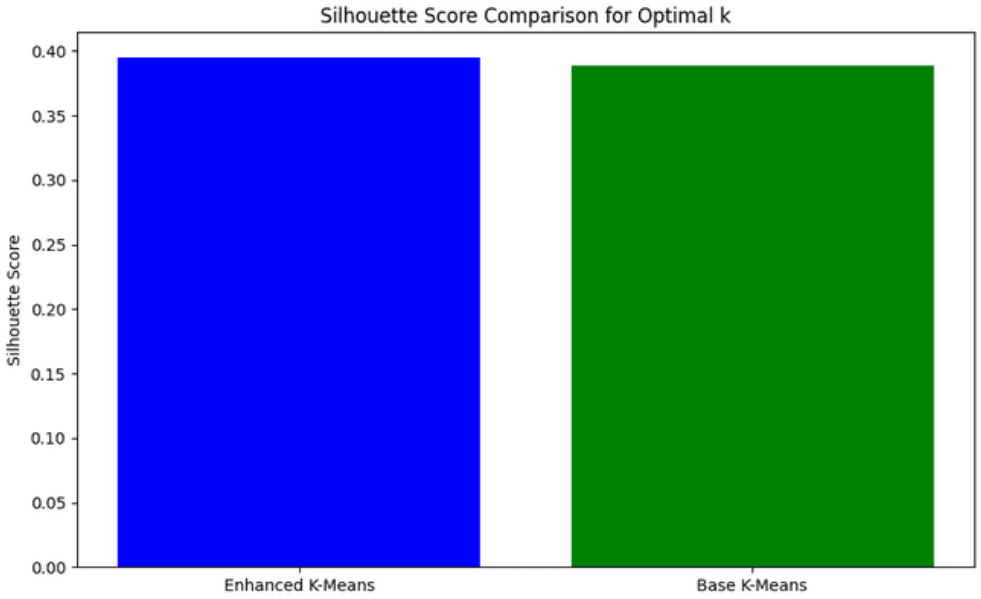
After handling the missing data, the dataset was scaled using StandardScaler to normalize the numeric columns. Scaling ensured that features were standardized, facilitating efficient, and accurate clustering.

4.2. Clustering Results

After the dataset was preprocessed, both the Base K-Means Algorithm and the Enhanced K-Means Algorithm were implemented to assess the impact of the proposed improvements. Specifically, the evaluation aimed to determine whether enhancements such as distance-weighted probability distribution initialization and mini-batch processing led to better clustering quality and greater computational efficiency than the standard approach. The algorithms' performance was compared based on two metrics discussed in 2.3: silhouette score and time complexity. Key results are discussed, including the optimal number of clusters, clustering quality, and computational efficiency.

Figure 2

Elbow method comparison of enhanced K-Means and base K-Means



The Elbow Method was employed to determine the optimal value for "k," which represents the number of clusters. The analysis revealed that k=4 was the optimal number of clusters for both algorithms, suggesting that earthquake occurrences naturally partition into four distinct groups. On the other hand, the Silhouette Score measures the quality of clustering by assessing how well data points align with their assigned clusters. Higher scores indicate better-defined clusters.

Figure 3

Silhouette score comparison

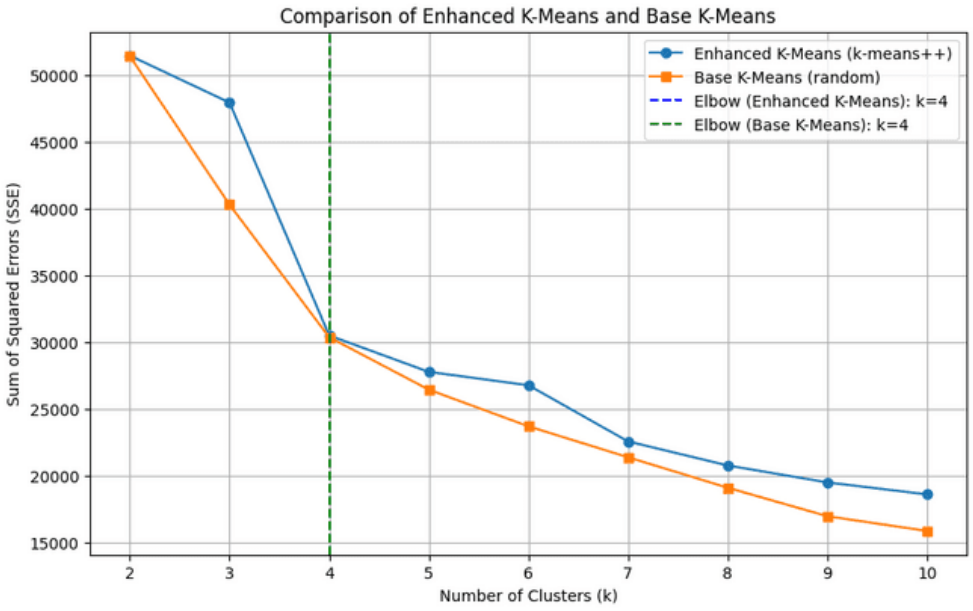


Table 1

Silhouette score comparison

	Score
Enhanced K-Means	0.3946
Base K-Means	0.3887

Table 2

Time complexity and number of iterations comparison

Algorithm	Iterations	Time Complexity (per run)	Total Time Complexity
Enhanced K-Means	1	600	6000
Base K-Means	10	838,760	8,387,600

While the difference is marginal, the enhanced K-Means (0.3946) outperformed base K-Means (0.3887), indicating better-defined clusters. While the improvement is relatively small, it aligns with the findings of Arthur and Vassilvitskii (2007), who introduced K-Means++ and demonstrated that improved centroid initialization significantly enhances clustering performance. Their study found that K-Means++ reduces the likelihood of poor centroid placement, which often leads to suboptimal clusters in standard K-Means. The enhanced K-Means algorithm in this study builds upon this principle by incorporating a distance-weighted probability distribution for centroid selection, further reducing randomness in initialization.

One of the key improvements introduced in the enhanced K-Means algorithm was the use of processing the dataset into mini batches, which significantly reduced computational effort. The enhanced K-Means converged after just 1 iteration, compared to 10 iterations required by the base K-Means. While for the time complexity result, the enhanced K-Means was 1,400 times faster per initialization run, making it substantially more efficient for large-scale datasets. This finding is consistent with Sculley (2010), who introduced mini-batch K-Means as a computationally efficient alternative to traditional K-Means for large datasets. The study demonstrated that processing data in smaller subsets significantly reduces computational time while maintaining clustering accuracy. Similarly, the integration of mini-batch processing allowed the enhanced K-Means algorithm to process the seismic data efficiently, making it a more practical choice for analyzing earthquake occurrence pattern clustering. These results highlight the computational advantage of enhanced K-Means,

particularly for large datasets where the base K-Means becomes impractical due to its high time complexity.

The enhanced K-Means algorithm demonstrated superior clustering performance and computational efficiency compared to the base K-Means algorithm, making it a valuable tool for seismic data analysis. By effectively preprocessing the dataset to handle missing values and standardize features, both algorithms identified k = 4 as the optimal number of clusters, indicating four natural groupings in the earthquake data. However, enhanced K-Means achieved better clustering consistency, as reflected in its higher Silhouette score, and significantly improved efficiency by converging in just one iteration. In contrast, base K-Means required multiple iterations, increasing computational costs.

With these results, the enhanced K-Means is recommended for large earthquake datasets due to its computational efficiency and clustering quality. These improvements make enhanced K-Means particularly applicable to large-scale seismic monitoring, where real-time processing is crucial for disaster response, risk assessment, and early warning systems. However, limitations exist, as the algorithm relies on a predefined number of clusters, which may not always capture complex seismic patterns. Additionally, the lack of spatiotemporal analysis prevents the algorithm from recognizing evolving earthquake trends. Future enhancements may further improve clustering outcomes and help identify underlying earthquake occurrence patterns by incorporating spatiotemporal analysis or applying feature engineering techniques.

5. Conclusion

The study introduced an enhanced K-Means algorithm, incorporating a distance-weighted initialization and mini-batch processing for clustering. The results demonstrate that the enhancements significantly improve clustering accuracy, computational efficiency, and scalability. These improvements result in a more precise and efficient clustering framework, enabling greater accuracy in the identification of high-risk seismic zones.

This research contributes to critical areas such as disaster preparedness, risk mitigation, and informed decision-making by offering a robust methodology for seismic data analysis. The enhanced algorithm supports better-informed decision-making in disaster preparedness and urban planning. Policymakers and disaster response teams can leverage this

tool to identify high-risk areas, tailor emergency interventions, and allocate resources more effectively, ultimately reducing the impact of seismic events on vulnerable populations.

Future researchers are encouraged to expand on this work by applying the enhanced K-Means algorithm to other geophysical datasets, including aftershock sequences, fault line proximity, or soil stability, to validate its robustness across diverse scenarios. Extending the algorithm for real-time clustering of streaming seismic data could provide immediate insights during seismic events, improving response times and disaster management strategies. These advancements would build on the foundation laid by this study, contributing to continued progress in earthquake clustering methodologies and global risk reduction efforts.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was not supported by any funding.

AI Declaration

The author declares the use of Artificial Intelligence (AI) in writing this paper. In particular, the author used ChatGPT in identifying relevant literature and refining content structure. The author takes full responsibility in ensuring that research idea, analysis and interpretations are original work.

References

Arthur, D., & Vassilvitskii, S. (2007). *k-means++: The Advantages of Careful Seeding*. https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf

Bottou, L., & Bousquet, O. (2008). The tradeoffs of large scale learning. *Advances in Neural Information Processing Systems* (NeurIPS), 20, 161–168.

Béjar, J. (2020). *K-means vs Mini Batch K-means: A comparison*. https://upcommons.upc.edu/bitstream/handle/2117/23414/R13-8.pdf

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, *40*(1), 200–210. https://doi.org/10.1016/j.eswa.2012.07.021

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., & Le, Q. V. (2012). Large Scale Distributed Deep Networks. *Advances in Neural Information Processing Systems* (NeurIPS), 25, 1223–1231.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD-96 Proceedings*. https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf

Fan, Z., & Xu, X. (2019). Application and visualization of typical clustering algorithms in seismic data analysis. *Procedia Computer Science*, *151*, 171–178. https://doi.org/10.1016/j.procs.2019.04.026

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (pp. 451–454). Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Springer series in statistics the elements of statistical learning data mining, inference, and prediction second edition*. https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf

Hicks, S. C., Liu, R., Ni, Y., Purdom, E., & Risso, D. (2021). mbkmeans: Fast clustering for single cell data using mini-batch k-means. *PLOS Computational Biology*, *17*(1), e1008625. https://doi.org/10.1371/journal.pcbi.1008625

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881–892. https://doi.org/10.1109/tpami.2002.1017616

Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition*, *36*(2), 451–461. https://doi.org/10.1016/s0031-3203(02)00060-2

Mato, F., & Theofilos Toulkeridis. (2017). An unsupervised K-means based clustering method for geophysical post-earthquake diagnosis. *IEEE Symposium Series on Computational Intelligence*, 1-8. https://doi.org/10.1109/ssci.2017.8285216

Novianti, P., Setyorini, D., & Rafflesia, U. (2017). K-Means cluster analysis in earthquake epicenter clustering. *International Journal of Advances in Intelligent Informatics*, *3*(2), 81. https://doi.org/10.26555/ijain.v3i2.100

Reynolds, D.A. (2009). Gaussian mixture models. In*:* Li, S.Z. and Jain, A., (Eds.), *Encyclopedia of Biometrics*. Springer. https://www.scirp.org/reference/referencespapers?referenceid=3466146

Rifa, I. H., Pratiwi, H., & Respatiwulan, R. (2020). Clustering of earthquake risk in Indonesia using K-Medoids and K-Means algorithms. *Media Statistika*, *13*(2), 194–205. https://doi.org/10.14710/medstat.13.2.194-205

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(0377-0427), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sanchez, M. Santibanez., Valdovinos, R. M., Trueba, A., Rendon, E., & Lopez, E. (2013). Applicability of cluster validation indexes for large data sets. *Artificial Intelligence (MICAI)*, 187–193. https://doi.org/10.1109/MICAI.2013.30

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

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. https://doi.org/10.1109/dsaa49011.2020.00096

Xiangyuan, H., Siyuan, L., & Hao, W. (2020). *A survey on k-means initialization methods*. https://www.dcs.warwick.ac.uk/~u2470130/randalg20/HLW.pdf

Xiao, B., Wang, Z., Liu, Q., & Liu, X. (2018). SMK-means: An improved mini batch K-means algorithm based on mapreduce with big data. *Cmc-Computers Materials & Continua*, *56*(3), 365–379. https://doi.org/10.3970/cmc.2018.01830

Xie, H., Zhang, L., Lim, C. P., Yu, Y., Liu, C., Liu, H., & Walters, J. (2019). Improving K-means clustering with enhanced Firefly Algorithms. *Applied Soft Computing*, *84*, 105763. https://doi.org/10.1016/j.asoc.2019.105763

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Xu, Y., Qu, W., Li, Z., Min, G., & Liu, Z. (2014). Efficient -Means++ Approximation with MapReduce. *IEEE Transactions on Parallel and Distributed Systems*, *25*(12), 3135–3144. https://doi.org/10.1109/TPDS.2014.2306193