Advanced Mathematical Models & Applications
Vol.10, No.1, 2025, pp.43-60
https://doi.org/10.62476/amma.10143

Jomard Publishing

A NOVEL GLOBAL OPTIMIZATION MODEL FOR n-DIMENSIONAL CENTER-BASED CLUSTERING PROBLEM AND ITS APPLICATION TO EARTHQUAKE INVESTIGATION

Ridwan Pandiya[1], **Atina Ahdika**[2*], **Siti Khomsah**[3]

[1]Department of Informatics, Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia
[2]Department of Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia
[3]Department of Data Science, Institut Teknologi Telkom Purwokerto, Banyumas, Indonesia

Abstract. Global optimization-based model for solving the data clustering problem depends on the type of the objective function. The existing model uses an exponential term in its objective function which apparently has a negative effect on the computational performance. This paper attempts to overcome the shortcomings experienced by the current model by initiating a new general form of the global optimization model. Computational performances of the proposed model is measured through the data clustering simulation. Three indicators, including CPU time, number of function evaluations, and number of iterations, are evaluated. Our new optimization model is then applied to earthquake investigation using Indonesian seismic data during 2022 and is compared to the existing model in order to evaluate the effectiveness and the predominance of our proposed model in discovering locations with various seismic activities.

Keywords: Center-based clustering, global optimization, optimal partition, earthquake, seismic activity.
AMS Subject Classification: 90C30, 90C26, 65K05.
Corresponding author: Atina Ahdika, Department of Statistics, Universitas Islam Indonesia, Jl. Kaliurang 14.5, Sleman, DIY Yogyakarta 55584, Indonesia, Tel.: +62274895920, e-mail: *atina.a@uii.ac.id*
Received: 1 October 2024; Revised: 16 October 2024; Accepted: 12 December 2024; Published: 11 April 2025.

1 Introduction

Unsupervised center-based clustering is a methodology to create groups of data having the same characteristics. Clustering procedure, which can be considered as an indirect data mining, has many benefits that have been implemented to solve problems in real life. Genome biology is one of field of studies that frequently applies clustering method. Results of experiments conducted by Genomicists on the genes of living things are usually expressed as expression arrays. A common strategy is to obtain genes that represent RNA which have similar characteristic, which can be studied more comprehensively in Kavitha & Tamilarasan (2020); Monti et al. (2003); Ray et al. (2007); Alok et al. (2020); Zhang et al. (2005). In the health sector, especially in psychology, clustering techniques are also extensively applied. This can be seen from the number of scholars in the field of psychology who conduct the research in this area. For instance, in Hodges & Wotring (2005), clinical diagnosis data, collected from various institutions providing mental health services, are used to identify the most impaired groups to the least impaired groups. The results of the study are very valuable in developing special programs, especially for groups

How to cite (APA): Pandiya, R., Ahdika, A., & Khomsah, S. (2025). A novel global optimization model for n-dimensional center-based clustering problem and its application to earthquake investigation. *Advanced Mathematical Models & Applications*, *10*(1), 43-60 https://doi.org/10.62476/amma.10143

most experiencing mental disorders. The availability of diagnostic data from various individuals, the support of very capable computer capabilities, and various cluster analysis methods, have made the exploitation of clustering algorithms very massive in the field of psychology. This has been summarized and can be studied further in Caroline et al. (2023). In market research, clustering techniques are also widely used, especially in determining market targets (see D'Urso et al. (2020); France & Ghose (2019); D'Urso et al. (2015)). Segmentation problems are also massively used in image segmentation. In this context, the image will be segmented into some layers (clusters) through the gray levels. The resulting optimum partition can be viewed as a compression of the original image. Some literature examining image segmentation using clustering techniques can be found in Zhu & Hunter (2019); Sharma & Ravinder (2023); Mittal et al. (2022). The last instance of the use of clustering techniques often employed is in the problem of identifying locations where earthquakes occur most frequently. The implementation of clustering procedures in the earthquake problem has a big contribution because the results can be used by stakeholders in making decisions in terms of preventive measures, especially for countries that frequently experience earthquakes such as Indonesia. Optimal partition in earthquake issues is usually based on location data, i.e. longitude, latitude, and magnitude. The authors in Scitovski & Scitovski (2013) attempted to seek both the local and global optimum partitions by proposing a new algorithm. The reliability of their algorithm especially in the term of the running time was then used as basis to observe the earthquake data in Croatia to determine the locations where seismic activity occurs most frequently. Seismic activity data will increase over time, so the opportunity for bias in clustering results will not be avoided. This problem was later captured by researchers in Seydoux et al. (2020) by offering a new framework based on the unsupervised deep learning. Their framework is able to cluster the continuous seismic data in Greenland landslide. Various methods and models of seismic data clustering can be explored further. For instance, subsequence time series clustering (Vijay & Nanda, 2023), spatio-temporal clustering (Yamagishi et al., 2021), and so forth.

In statistical learning theory (better known as machine learning), clustering methods fall into the unsupervised learning category. Methods for solving clustering problems are extensive. There are at least 4 taxonomies in clustering algorithms. From the technique of the algorithm point of view, clustering methods can be divided into centre-based partitioning clustering, hierarchical clustering, density-based clustering, and model-based clustering. Partitional clustering divides data into several clusters where the number of clusters has been determined previously. Clusters are organized based on certain criteria or based on their similarities. Hierarchical clustering deals with organizing data into a sequence of nested data partitions in a hierarchical structure. Different from the previous two traditional clustering, density-based clustering enables clusters with various sizes and shapes. The last type of clustering, model-based clustering, allows the formed data distribution to be supported by latent subsets.

From the clustering results point of view, clustering algorithm can also be categorized as soft and hard clustering. In hard clustering, all the data between clusters are well separated, whereas soft clustering, can also be referred as fuzzy clustering, allows for data intersection. In other words, one data can fall into several clusters (Gao et al., 2023). This study focuses in hard and center-based partitional clustering. How the data is grouped based on each clusters requires measuring instruments. The most frequently used is the concept of distance. In center-based clustering for continuous variables, for instance, the concept of Euclidean norm is commonly used. However, Euclidean distance is not the only distance measure. Generally, the distance function will be adjusted to the type of data and the outlier data problem. It was later discovered that the Euclidean distance is irritable to outliers (Jain et al., 1999). To anticipate sensitivity to outliers such as those that occur due to the use of Euclid distance, other distance concepts are used, namely the Manhattan distance and Chebyshev distance. Those distances use the absolute value of the difference between data and its center instead of implementing the square term like Euclidean norm do. Therefore, the Manhattan and chebyshev distance are more robust

to outliers (Rui & Wunsch, 2005). However, among the many types of distance measures, the Mahalanobis distance is considered to best accommodate the problem of data outliers. This is because this distance takes into account data variations and data correlations. However, the use of the variance-covariance matrix in its calculations leaves serious computational problems (De Maesschalck et al., 2000). In real-world applications, data can vary widely. If the data we are dealing with is not numerical data, for example the frequency of occurrence of words in a document, it will be quite difficult if it has to be converted into numerical data. Therefore, geometric distance measures become less relevant. To overcome this issue, the cosine distance or dot product are used so that bias, especially in text mining, can be avoided (Tang et al., 2023). Apart from the diversity of data types, sometimes we are also faced a very wide range of data values for each category. Therefore the data is normalized. Euclidean distance is not suitable to be used because it will cause a "double-zero" effect, which means that the number 0 does not represent similarity. To avoid the intended bad effects, Chord distance is usually implemented (Ricotta, 2019).

Distance-based (or well known as center-based) algorithms in solving clustering problems use one of the sparse measures mentioned previously. One of the most representative of the center-based clustering methods is the K-means algorithm. The main concept of the K-means algorithm is to determine mutually exclusive clusters based on Euclidean distance of the data points to the cluster center with the cluster shape being is a sphere. This algorithm was firstly found in the 1950s. Its massive use for various real problems until recently is a proof that the K-means method is the most popular data clustering method. However, it does not mean that the K-means algorithm is without limitations. Among the many advantages of this algorithm, for example in terms of computing speed, the K-means algorithm still has several disadvantages. There are at least 3 limitations that are often experienced when applying this algorithm. Studies related to big data inevitably deal with outlier data. Unfortunately, the K-means method is considered too sensitive to outlier data. The second drawback is that the K-means method cannot accommodate non-spherical cluster shapes. The last shortcoming is that the K-means algorithm sometimes can only gain a local optimum solution. This is mainly due to the involvement of random starting points in the initial stage of this algorithm (Jain, 2010). The improvement of the initial starting point issue is then fixed by the K-means++ method, while K-medoids algorithm participates in terms of increasing its computational speed. CLARA and CLARANS methods were proposed to simplify the K-medoids algorithm when used for clustering very large data (see Schubert & Rousseeuw (2021)). As we previously mentioned that the Euclidean distance behaves less efficiently when dealing with outlier data. Following up on this issue, the K-medians method replaces this distance measure with the Manhattan distance. A type of K-means method specifically for categorical data are K-modes and K-prototypes algorithms, while the problem of overlapping clusters and non-spherical cluster types can be overcome using the fuzzy C-means and Kernel K-means algorithm, that can be studied more comprehensively in Hashemi et al. (2023) and Guan & Terada (2023). Although this paper focuses on the center-based clustering problem, readers who are willing to study some methods considered as hierarchical clustering problems can read in more detail in Mullner (2013); Guo et al. (2021); Xie et al. (2023); Zhang et al. (1997); Liang & Chen (2021); Guha et al. (2000) and the non-sphere shaped clustering can be learned in Wang et al. (2021); Wu et al. (2023); Oyewole & Thopil (2023); Latifi-Pakdehi & Daneshpour (2021); Yuan et al. (2021). Finally, model-based clustering approaches have been widely researched and published. Some reference can be found in Furno (2023); Weber et al. (2022); Bandeen-Roche et al. (1997).

All of the center-based clustering methods mentioned focus on developing algorithms that use distance functions. The sum of the distances of data points to its center acts as an objective function that must be minimized. From the optimization point of view, the key success of the process of obtaining the minimum point is influenced by the type of objective function to be minimized and the type of method used to minimize the objective function. For instance, a

continuously differentiable objective function will be easier to be minimize compared to a non-differentiable objective function. From an algorithmic perspective, metaheuristic methods are more applicable compared to deterministic methods, although metaheuristic methods do not have the guarantee of convergence that deterministic methods have. The contribution of this research is to formulate a new objective function where the minimum points of the objective function are the centers of the data to be found. Our objective function formulation process is motivated by an optimization model previously developed in Sabo et al. (2013). The authors in Sabo et al. (2013) proposed an exponential-based optimization model for one-dimensional clustering problem. From our analytical and computational experiments, there are at least two drawbacks suffered by the clustering model (Sabo et al., 2013). First, it includes a single parameter that difficult to be accommodated. Second, the exponential function included in the optimization model offered by Sabo et al. (2013) can be possibly caused an expensive computational process. Several attempts were made by the author in Scitovski (2017) by reducing the global optimization model to find optimal clusters into a hypercube $[0,1]^k$, where k is the number of clusters. Not only that, to reduce the complexity, the optimization model is transformed by using an appropriate linear transformation. However, how to find a right linear transformation is not an easy problem. This paper aims to propose an optimization model with a non-exponential and non-logarithmic functions. The superiority and effectiveness of the proposed model is demonstrated by numerical simulations. The proposed optimization model is also implemented to earthquake inspection using Indonesian seismic data during 2022 concerning to group locations with various seismic activities.

This paper is mapped into six sections. Section 2 is devoted to presenting several definitions, concepts, assumptions, and notations that will be used throughout this paper. A complete explanation of the proposed optimization model will be presented in Section 3. Section 4 contains the data used in the numerical simulation process, the minimization method, and the results of the numerical simulation. Application to the real earthquake investigation is given in Section 5 along with its comparison to the existing model and the conclusion is drawn in Section 6.

2 Optimization-based data clustering

This study aims to solve the n-dimensional data (data which have the n features) clustering problems, as can be illustrated in Table 1. For each row of Table 1 is considered as a point. Therefore, Table 1 contains x_i data, where $i = 1, \ldots, k$, and $x_i \in \Omega \subseteq R^n$. For instance, the third point of data given in Table 1 is $x_3 = \left(a_3^1, \ldots, a_3^n\right)$.

Table 1: The illustration of n-dimensional data

feature 1	feature 2	feature 3	feature 4	$\cdots$	feature n
a_1^1	a_1^2	a_1^3	a_1^4	$\cdots$	a_1^n
a_2^1	a_2^2	a_2^3	a_2^4	$\cdots$	a_2^n
a_3^1	a_3^2	a_3^3	a_3^4	$\cdots$	a_3^n
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k^1	a_k^2	a_k^3	a_k^4	$\vdots$	a_k^n

The problem is to group the data into j subgroup (partition) $\Phi = \{\beta_1, \ldots, \beta_j\}$, in which all the clusters in the partition Φ are well separated ($\bigcap_{m=1}^{j} \beta_i \neq \emptyset$). The problem will be obtained by completing the following algorithm.

1. Find a center α_m for every clusters β_m in Φ, with $m = 1, \ldots, j$.

2. Calculate the sum of the distances from elements of β_m to α_m $\left(\sum_{x_i \in \beta_m} \chi(\alpha_m, x_i)\right)$, where χ is the distance function.
3. Minimize $\Psi(\beta_m) = \sum_{x_i \in \beta_m} \chi(\alpha_m, x_i)$.

The distance $\chi(.,.)$ has many types. Therefore, we can choose one that fits Definition 1. One of the most frequently used is $\chi(y_1, y_2) = \|y_1 - y_2\|^2$ (least squares function).

Definition 1. *A real valued function $\chi : R^n \times R^n \to R_+$, where $R_+ = [0, +\infty)$, is called a distance-like function if it is satisfied the following three properties.*

1. *$\chi(y_1, y_2) = 0$ if and only if $y_1 = y_2$*
2. *$y_1 \mapsto \chi(y_1, y_2)$ is continuous on R^n, $\forall y_2 \in R^n$*
3. *χ is globally convex, i.e. $\lim_{\|y_1\| \to +\infty} \chi(y_1, y_2) = +\infty, \forall y_2 \in R^n$.*

From the previous explanation, it is clear that we seek the following global optimal partition.

$$\text{Minimize } \Psi(\Phi), \tag{1}$$

where Φ is the partition, and $\Psi(\Phi) = \sum_{m=1}^{j} \sum_{x \in \beta_m} \chi(\alpha_m, x)$. However, from the global optimization theory, partition Problem (1) cannot be solved since Φ is a partition, not variables. Hence, one can reformulate Problem (1), into Problem (2)

$$\underset{\alpha \in R^{jn}}{\text{minimize}}\ P(\alpha), \tag{2}$$

where $P : R^{jn} \to R_+$ and $P(\alpha) = \sum_{i=1}^{n} \min_{1 \leqslant m \leqslant j} \chi(\alpha_m, x_i)$. Problem (2) is non-convex, multimodal, nonlinear, symmetric, Lipschitz-continuous, and could be a non-differentiable optimization problem. However, many studies have shown that continuously differentiable functions are easier to minimize by implementing traditional minimization procedure. Based on this fact, Teboulle (2007); Iyigun & Ben-Israel (2010); Sabo et al. (2013) offered the smoothing technique by reformulating Problem (2) into continuously differentiable global optimization model defined as

$$\min_{\alpha_m \in R^n} P(\alpha_1, \ldots, \alpha_j), \quad P(\alpha_1, \ldots, \alpha_j) = -\varepsilon \sum_{i=1}^{n} \ln \sum \exp\left(-\frac{\|\alpha_m - x_i\|^2}{\varepsilon}\right), \tag{3}$$

where ε is a small positive real number and $\|.\|^2$ is a least squares distance-like function. There are several global optimization techniques to solve Problem (3). Some nature-based approached, including the genetic algorithm (Maulik & Bandyopadhyay, 2000), particle swarm optimization (PSO) (Rengasamy & Murugesan, 2021), whale optimization algorithm (Nasiri & Khiyabani, 2018), simulated annealing algorithm (Selim & Alsultan, 1991), and bee colony algorithm (Rahnema & Gharehchopogh, 2020).

From the theoretical optimization perspective, Problem (3) can be solved easily because $P(\alpha_1, \ldots, \alpha_j)$ is continuously differentiable. The continuously differentiable property implies that there will be many efficient gradient-based minimization procedures that can be applied Liu et al. (2017). However, the exponential and logarithmic functions involved in $P(\alpha_1, \ldots, \alpha_j)$ can lead to the overflow effect. The effect will occur especially when the value of $\|\alpha_m - x_i\|^2$ is large or the value of the parameter ε is small. Moreover, the overflow phenomenon will affect the

computational performances of (3), which is usually characterized by an increase in the running time, number of iterations, and also a significant increase in the number of function evaluations. This situation makes the minimization process of Problem (3) inefficient. This research will fill this research gap by proposing a new type of objective function that can overcome the weaknesses of Problem (3), which will be discussed in depth in Section 3.

3 A new optimization model for data clustering

The previous explanation regarding the optimization perspective in the context of data clustering leads us to a dilemma. If the model is left without a smoothing process (changing Problem (2) into Problem (3) without involving exponential and logarithmic terms), then the clustering problem will become a black box optimization problem. Solving this type of problem is by no means an easy matter. On the other hand, if the smoothing process involves exponential and logarithmic functions, it tends to be easier to complete but will have an expensive impact on computational process. Based on the mathematical analysis point of view, we have a hypothesis that if the exponential and logarithmic functions in Problem (3) are replaced with other functions with similar properties, then numerical computation will not be too difficult, especially when solving clustering problems with very large volumes of data. Since functions which have similar properties to the exponential and logarithmic functions are not unique, there will be many functions to choose from. Therefore, the proposed objective function will be in a general form, which is done through the following several mechanisms:

1. Objective function P in Problem (3) is broken down into two functions $\zeta_1 : \mathbb{R} \to \mathbb{R}$ and $\zeta_2 : \mathbb{R} \to \mathbb{R}$, where $\zeta_1(s) = \exp\left(-s^2\right)$ and $\zeta_2(r) = \ln(r)$. Since ε is a fixed small real number, then we can rule it out.
2. Function ζ_1 operates on $[0, \infty)$ because $\|\alpha_m - x_i\|^2$ is always non-negative, while ζ_2 has the domain $(0, \infty)$ due to the fact that $\exp\left(-\|\alpha_m - x_i\|^2\right)$ is always positive.
3. The next step is to identify the analytical and geometrical properties of functions ζ_1 and ζ_2.
4. The results of the properties of the functions ζ_1 and ζ_2 that have been obtained from stage 3 are used as a reference for finding other functions for both ζ_1 and ζ_2 which will be used as substitutes for exponent and logarithm function in the objective function of Problem (3).

Based on the described mechanisms, we propose the following new general form of the optimization model for solving the n-dimensional data clustering

$$\min_{\alpha_m \in R^n} P(\alpha_1, \ldots, \alpha_j), \; P(\alpha_1, \ldots, \alpha_j) = -\varepsilon \sum_{i=1}^{n} \zeta_1 \sum_{m=1}^{j} \zeta_2 \left(-\frac{\|\alpha_m - x_i\|}{\varepsilon} \right), \tag{4}$$

where ζ_1 and ζ_2 are real valued functions, which have the following properties:

1. ζ_1 and ζ_2 are continuously differentiable on $\mathbb{R}$
2. $\zeta_2(s) > 0$ on $\mathbb{R}$
3. $\zeta_2(0) = c$, where c is a positive real number
4. $\zeta_2(s) \searrow 0$ as $|s| \to \infty$
5. $\zeta_2'(s) \geqslant 0$ for every $s \in (-\infty, 0]$ and $\zeta_2'(s) = 0$ if and only if $s = 0$, where $\zeta_2'(s)$ is the derivative of ζ_2 at s

6. $\zeta_1(1) = 0$
7. $\zeta_1(r)$ is strictly decreasing on $(0, \infty)$
8. $\zeta_1(r) > 0$ on $(0,1)$ and $\zeta_1(r) < 0$ on $(1, \infty)$
9. $\zeta_1'(r) < 0$, $\forall r \in (0, \infty)$
10. $\zeta_1'(r)$ is strictly increasing on $(0, \infty)$
11. $\zeta_1'(r) \nearrow 0$ as $r \to \infty$.

Some examples of functions ζ_1 include $\zeta_1(r) = -\arccos\left(-\frac{r}{1+r}\right) + \frac{2\pi}{3}$, $\zeta_1(r) = \arctan(-r) + \frac{\pi}{4}$, $\zeta_1(r) = \arcsin\left(-\frac{r}{1+r}\right) + \frac{\pi}{6}$, $\zeta_1(r) = \operatorname{arc\,cot}(r) - \frac{\pi}{4}$, $\zeta_1(r) = \frac{1}{1+r^2} - \frac{1}{2}$, while some examples of ζ_2 are $\zeta_2(s) = \frac{1}{1+s^2}$, $\zeta_2(s) = -arc\cot\left(-s^2\right) + \pi$, $\zeta_2(s) = \left(s^2+1\right)^{-\frac{1}{2}}$, $\zeta_2(s) = \arctan\left(-s^2\right) + \frac{\pi}{2}$, $\zeta_2(s) = \arcsin\left(-\frac{s^2}{1+s^2}\right) + \frac{\pi}{2}$, $\zeta_2(s) = \tanh\left(-s^2\right) + 1$.

Up to this point, an optimization model for clustering high-dimensional data (Problem 4) has been provided. The next problem is to show that the proposed model is executable. First, it is most important to point out that the objective function in (4) is bounded below, which is demonstrated by Theorem 1.

Theorem 1. *If $A = \{x_i \in R^n : i = 1, \ldots, k\} \subset \Omega$, where $\Omega = \left\{a \in R^n : l^i \leqslant a^i \leqslant u^i\right\}$ for some $l = \left(l^1, \ldots, l^n\right)$ and $u = \left(u^1, \ldots, u^n\right)$, then the objective function*

$$P_\varepsilon(\alpha) = -\varepsilon \sum_{i=1}^{n} \zeta_1 \sum_{m=1}^{j} \zeta_2\left(-\frac{\|\alpha_m - x_i\|}{\varepsilon}\right)$$

is bounded below.

Proof. Let

$$P(\alpha) = \sum_{i=1}^{n} \min_{m=1,\ldots,j} \|\alpha_m - x_i\|^2$$

and

$$P_\varepsilon(\alpha) = -\varepsilon \sum_{i=1}^{n} \zeta_1 \sum_{m=1}^{j} \zeta_2\left(-\frac{\|\alpha_m - x_i\|}{\varepsilon}\right).$$

To proof that the objective function in (4) is bounded below, we need to consider the following difference between $P(\alpha)$ and $P_\varepsilon(\alpha)$

$$P(\alpha) - P_\varepsilon(\alpha) = \sum_{i=1}^{n} \min_{m=1,\ldots,j} \|\alpha_m - x_i\|^2 + \varepsilon \sum_{i=1}^{n} \zeta_1 \sum_{m=1}^{j} \zeta_2\left(-\frac{\|\alpha_m - x_i\|}{\varepsilon}\right).$$

For simplicity, $\frac{\|\alpha_m - x_i\|}{\varepsilon}$ will be denoted as Γ_{mi}. Furthermore, we will assume (without lost of generality) that $\Gamma_{m1} \leq \ldots \leq \Gamma_{mi}$. Therefore, we have

$$P(\alpha) - P_\varepsilon(\alpha) = \sum_{i=1}^{n} \min_{m=1,\ldots,j} \varepsilon\Gamma_{mi} + \varepsilon \sum_{i=1}^{n} \zeta_1 \sum_{m=1}^{j} \zeta_2(-\Gamma_{mi}).$$

Since $\Gamma_{m1} \leq \ldots \leq \Gamma_{mi}$, thus, $\min\limits_{m=1,\ldots,j} \Gamma_{mi} = \Gamma_{m1}$. It follows that

$$P(\alpha) - P_\varepsilon(\alpha) = \varepsilon \sum_{i=1}^{n} \Gamma_{m1} + \varepsilon \sum_{i=1}^{n} \zeta_1 \sum_{m=1}^{j} \zeta_2(-\Gamma_{mi}) = \varepsilon \sum_{i=1}^{n} \left(\Gamma_{m1} + \zeta_1 \sum_{m=1}^{j} \zeta_2(-\Gamma_{mi})\right).$$

We know that Γ_{mi} is positive. From the property of the norm, it implies that

$$0 \leq \Gamma_{mi} - \Gamma_{m1} = \frac{\|\alpha_m - x_i\|}{\varepsilon} - \frac{\|\alpha_m - x_1\|}{\varepsilon} \leq \frac{1}{\varepsilon} \|\alpha_m - x_i\| < \frac{1}{\varepsilon} (u - l).$$

Hence, the following inequality holds

$$\varepsilon \sum_{i=1}^{n} \left(\Gamma_{m1} + \zeta_1 \sum_{m=1}^{j} \zeta_2 (-\Gamma_{mi}) \right) \geq \varepsilon \sum_{i=1}^{n} \left(\zeta_1 \sum_{m=1}^{j} \zeta_2 (-\Gamma_{mi}) \right).$$

From properties (3), (4), and (5) of ζ_2, it is clear that ζ_2 attains its maximum value at c. It is implies that

$$\varepsilon \sum_{i=1}^{n} \left(\zeta_1 \sum_{m=1}^{j} \zeta_2 (-\Gamma_{mi}) \right) \geq \varepsilon n \zeta_1 ((j-1)c),$$

where c is a positive real number as mentioned in the property (4) of ζ_2. On the other hand, since ζ_1 is strictly decreasing on $(0, \infty)$, then $\zeta_1 \left(\sum_{m=1}^{j} \zeta_2 (-\Gamma_{mi}) \right) < j$. Therefore, it follows that

$$\varepsilon \sum_{i=1}^{n} \left(\Gamma_{m1} + \zeta_1 \sum_{m=1}^{j} \zeta_2 (-\Gamma_{mi}) \right) < \frac{1}{\varepsilon} n (u - l) j.$$

Thus,

$$\varepsilon n \zeta_1 ((j-1)c) \leq P(\alpha) - P_\varepsilon(\alpha) < \frac{1}{\varepsilon} n (u - l) j,$$

or we can write it as

$$P_\varepsilon(\alpha) > P(\alpha) - \frac{1}{\varepsilon} n (u - l) j > -\frac{1}{\varepsilon} n (u - l) j.$$

In conclusion, $P_\varepsilon(\alpha)$ is bounded below. □

Theorem (1) has proven that the objective function in (4) is bounded below. Moreover, from the assumption of Theorem (1), Ω is closed and bounded set. Therefore, the existence of the global minima of $P_\varepsilon(\alpha)$ guaranteed.

4 Computational Simulation

In the previous section we have given a new general form of optimization model intended for use in data clustering problems. Some specific examples of the general forms conceived have also been given. Of course, in the implementation phase, we only need to choose one specific form from the general form of the model. The choice of specific form can be adjusted to the type of data. It is possible that for the same data the computational efficiency will be different if different specific forms are applied. This section is designed to show that the proposed model can be used to find the most optimal centroid. The specific forms we have chosen for the purposes of numerical experiments in this section are as follows:

$$P(\alpha_1, \ldots, \alpha_j) = -\varepsilon \sum_{i=1}^{m} \arctan \left(\sum_{m=1}^{j} \arctan \left(-\frac{\|\alpha_m - x_i\|^2}{\varepsilon} \right) \right) + \frac{\pi}{4}.$$

The experiment conducted is designed using some mechanism, including determining the numerical optimization method, data preparation, and so on. All the mechanism and its explanation and the computational results will be given in the following subsections.

4.1 Determine the Global Minimization Method

In almost all literature, the optimization model is intended only to find a starting centroid when implementing the K-means algorithm. However, in principle, the optimization problem (4) can be directly solved using any global minimization method, whether the method is included in the stochastic or deterministic approach. In the experiments carried out in this research, we took the option of solving the clustering problem by directly executing the optimization model without using the K-means algorithm. Both stochastic and deterministic approaches have their positive and negative aspects.

The stochastic or metaheuristic approach often deals with convergence that has no guarantee analytically, but on the other hand it is very applicable because carrying out its methods does not need to involve many assumptions, even if the objective functions do not have a closed form. On the other hand, deterministic approach is very promising in theory but it is poor when implemented in real problems, especially because deterministic methods require many assumptions. However, thanks to the authors in Jones et al. (1993) (and other DIRECT method developers which can be seen in Costa et al. (2018); Gablonsky & Kelley (2001); Lovison & Miettinen (2021)) for paving the way for the tension between the two types of approaches. In Jones et al. (1993), it was introduced the method well known as the DIRECT method. Unlike deterministic methods, which usually require a starting point, the DIRECT method does not require this, instead, DIRECT method use sample points determined by a certain formula. For some scholars the DIRECT method is also called a sample-based deterministic method. This fact is in line with the hole that needs to be filled related to the shortcomings of the k-means method relating to the starting point of the centroid which often produces different centroids. Those rationales motivate us to implement the DIRECT algorithm to solve Problem (4).

4.2 Data Preparation

Since the purpose of this section is to test whether the proposed model can be applied in partitioning data and then if it is reliable, the computational performance will be tested, the data used in the numerical experiment section is randomly constructed data. We use three types of data volume, namely 1000 data, 5000 data, and 10000 data, on the interval $[1, 100]$. For each type of data, 2, 3, 4 and 5 clusters will be determined.

4.3 Code Preparation

To solve the clustering problem using (4), we need to write some program codes. In our experiments, we use Matlab works in Windows 11 with a processor Intel(R) Core(TM) i7-9700 CPU @ 3.00GHz (8 CPUs), 16384MB RAM. Program codes are divided into 3 parts, they are the procedure of the objective function display in (4), procedure of the DIRECT method, and the last one is the main program to execute the clustering problem and to display the computational results. For the DIRECT method, since it has been widely developed even up to nowadays, by considering its advantages and suitability for the problem solved in this research, we chose to employ the modified DIRECT method carried out by authors in Bjorkman & Holmstrom (1999).

4.4 Computational Results

In this section, we execute the objective function in (4) by using the DIRECT method. The value of ε in (4) is in the interval $[0, 1]$. During the computational stages, some indicators are

examine, includes

$\mathcal{N}_D$: the number of data,
$\mathcal{N}_F$: the number of features (dimension of data),
$\mathcal{N}_C$: the number of clusters
ε : The value of the parameter ε,
$\mathcal{F}_E$: The number of function evaluations,
$\mathcal{R}_T$: running time until the algorithm stops
$\mathcal{N}_I$: the number of iterations to obtain the global optimum centroid

Table 2: The computational results

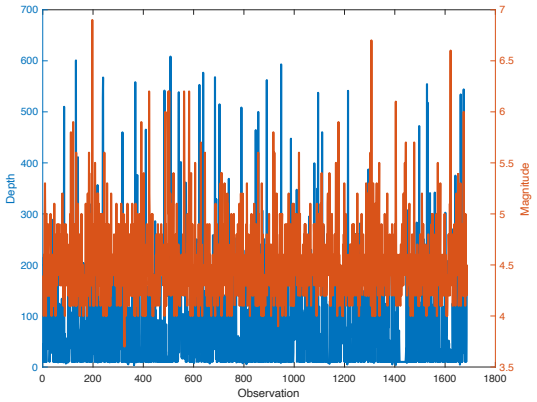
$\mathcal{N}_D$	$\mathcal{N}_F$	$\mathcal{N}_C$	ε	$\mathcal{F}_E$	$\mathcal{R}_T$	$\mathcal{N}_I$
1000	2	2	1.000	349	30.812	22
1000	2	3	1.000	423	34.984	17
1000	2	4	0.100	1071	120.936	34
1000	2	5	1.000	291	31.905	7
1000	3	2	0.100	67	21.948	4
1000	3	3	0.100	83	9.239	3
1000	3	4	0.010	1503	468.477	30
1000	3	5	0.010	1403	310.965	20
1000	5	2	0.010	665	112.014	13
1000	5	3	0.010	441	43.488	6
1000	5	4	0.010	3205	991.45	29
1000	5	5	0.010	3489	1085.23	35
5000	2	2	0.001	41	9.231	5
5000	2	3	0.001	261	46.302	13
5000	2	4	0.001	677	164.342	19
5000	2	5	0.001	811	220.293	18
5000	3	2	0.001	361	68.126	17
5000	3	3	0.001	625	161.29	15
5000	3	4	0.001	829	237.581	14
5000	3	5	0.001	1291	439.603	17
5000	5	2	0.001	565	143.284	10
5000	5	3	0.001	1503	868.25	15
5000	5	4	0.001	2147	1097.635	20
5000	5	5	0.001	2749	1701.237	19
10000	2	2	0.001	15	8.893	2
10000	2	3	0.001	23	13.264	2
10000	2	4	0.001	57	28.657	3
10000	2	5	0.001	73	37.419	3
10000	3	2	0.010	201	73.269	10
10000	3	3	1.000	661	270.768	15
10000	3	4	0.001	1163	743.837	22
10000	3	5	0.001	2135	1652.543	28
10000	5	2	0.001	713	294.01	13
10000	5	3	0.001	1695	1461.254	17
10000	5	4	0.001	3013	1965.086	20
10000	5	5	0.001	4261	3276.766	24

The computational results are displayed in Table 2. Table 2 shows that the optimization model proposed in this paper can solve clustering problems for data that has more than one dimension. The computing performance appears to be efficient, especially from the perspective of the number of function evaluations and the number of iterations. CPU time is of course greatly influenced by the device employed. In general, it can be concluded that the greater the number of partitions to be formed, the greater the number of iterations and function evaluations, which of course is also linear in the length of algorithm execution. The reliability and the competitiveness compared to other method of the optimization model proposed in this paper will also be tested

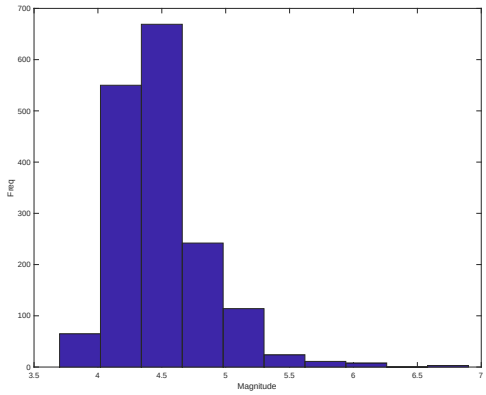
to solve real problems, namely earthquake affected clustering problem, which will be discussed in the next section.

5 Application to earthquake data

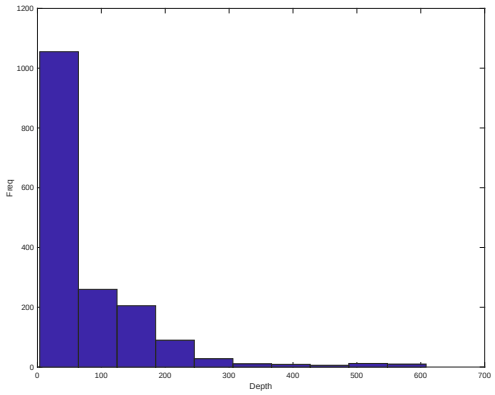
In this section, we utilize our proposed model to cluster the earthquake affected areas in Indonesia in 2022 based on two main earthquake measures, i.e. the depth and the magnitude of the earthquake. The earthquake data consists of 1686 points spread throughout Indonesia with a magnitude between 3.7 and 6.9 on the Richter scale and a depth between 3.604 and 608.287 km. Fig. 1 shows the distribution of the pair of earthquake magnitude and depth data from all observations.



(a) Earthquake observations



(b) Histogram of earthquake magnitude



(c) Histogram of earthquake depth

Figure 1: Earthquake data in Indonesia in 2022

Based on Fig. 1, the average values of earthquake magnitude and depth are 4.501 Richter scale and 76.121 km, respectively. These two average values will later become a reference in discussing the clustering results. Fig. 1a shows that, in general, the magnitude of the 1686 earthquake points were spread almost closely as it has close values. However, some points have a greater value than the average and very few points have smaller one. When correlated with the histogram in Fig. 1b, the distribution of earthquake magnitudes appears to be approximately normal. Consequently, when clustering is performed, the cluster centers are likely to be around the mean values. In contrast, the distribution of earthquake depth data (see Fig.1c) shows a

higher frequency at lower depths and tends to be right-skewed. This suggests that the cluster centers may have relatively similar values or even significantly different ones.

To evaluate the performance of our proposed model, in this application section, we compare our clustering method with its predecessor, the Sabo model (Sabo et al., 2013). The clustering results for two and three clusters can be seen in Table 3.

Table 3: Clustering results for earthquake data in Indonesia in 2022

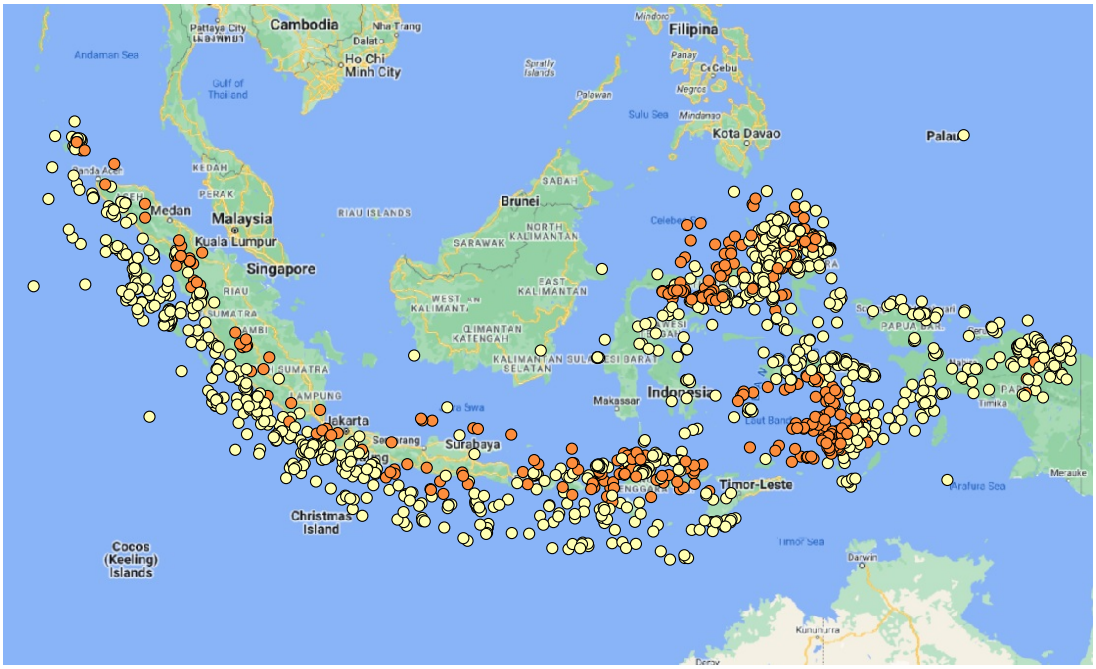
2 Clusters							
Model	Cluster	Centroid		ε	$\mathcal{F}_E$	$\mathcal{R}_T$	$\mathcal{N}_I$
		Depth	Magnitude				
Sabo	1	26.414	4.489	0.05	2747	269.787	37
	2	160.788	4.334				
Proposed	1	52.128	4.466	0.05	941	185.592	23
	2	52.128	4.438				
3 Clusters							
Model	Cluster	Centroid		ε	$\mathcal{F}_E$	$\mathcal{R}_T$	$\mathcal{N}_I$
		Depth	Magnitude				
Sabo	1	22.267	4.381	0.05	11767	631.478	50
	2	37.198	4.996				
	3	164.106	4.325				
Proposed	1	52.128	4.438	0.05	7185	392.608	49
	2	59.593	4.438				
	3	44.663	4.466				

We compare the performance of our proposed method with the Sabo method by evaluating several performance metrics, including the number of function evaluations $\mathcal{F}_E$, running time until the algorithm stops $\mathcal{R}_T$, and the number of iterations to obtain the global optimum centroid $\mathcal{N}_I$. Based on the results presented in the Table 3, it can be seen that for all three performance metrics, our proposed model demonstrates superior performance, as indicated by the smaller values of the performance metrics. Although the difference in the number of iterations required to reach the global optimum is not very significant, our model is much more effective in terms of running time, reducing computation time by approximately 60 to 70 percent compared to the Sabo model. Additionally, in terms of the number of function evaluations, our model requires approximately 30 to 65 percent fewer evaluations than the Sabo model.

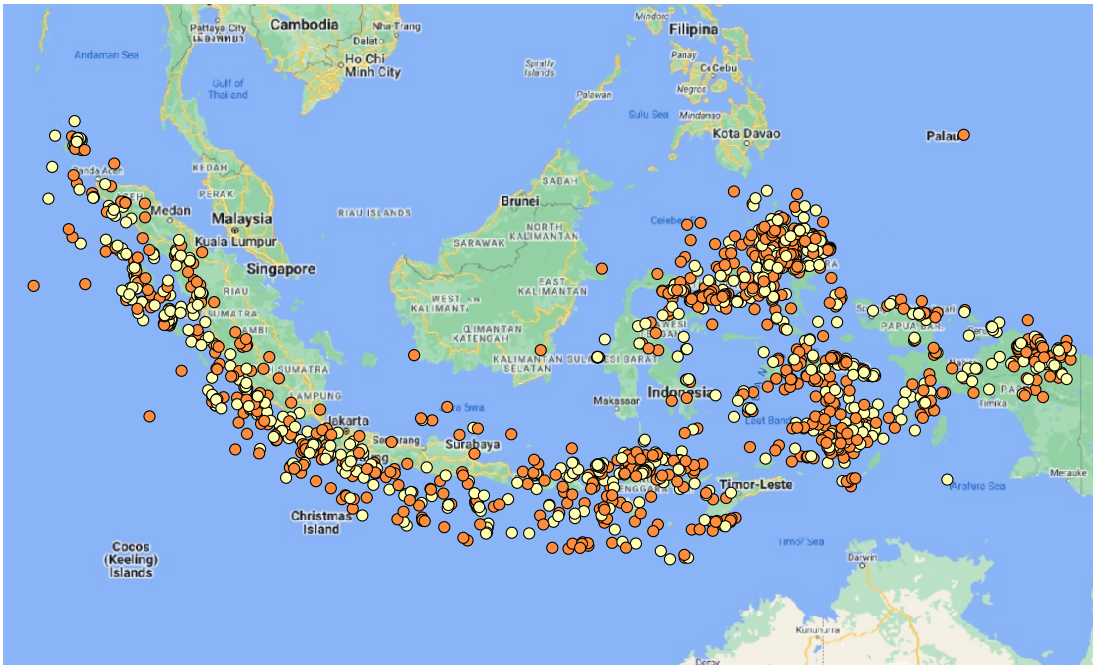
To gain a deeper understanding regarding the Sabo model and our proposed model in both clustering experiments, we created a mapping map of the clustering results presented in Figs. 2 and 3.

We can see how the clusters are distributed over the Indonesian region in the clustering result that was produced using the Sabo model. The groups are dispersed across the entire archipelago, encompassing Papua, Java, Kalimantan, Sulawesi, and Sumatra. Larger, more broadly based clusters that span wide geographic regions likely to arise in the Sabo model clusters. This leads to clusters that may include areas with different amounts of seismic activity.

Our proposed model's clustering result also reveals a widely dispersed cluster distribution across Indonesia. Nonetheless, the proposed model's clusters seem more granularly fine-grained and more locally concentrated than the Sabo model's. This indicates that the suggested model can more accurately locate and distinguish between areas with various seismic properties. The clustering centers are more densely packed, reflecting a more detailed analysis of the earthquake data. A higher density of the cluster centers indicates a more refined and localized clustering approach, this could be beneficial for targeted earthquake preparedness and mitigation strategies. Furthermore, for more detail information, we have made the clustering profile for each experiment consists of the number of member in each cluster, and the average values of the earthquake depth and magnitude for each cluster, which is provided in Table 4.

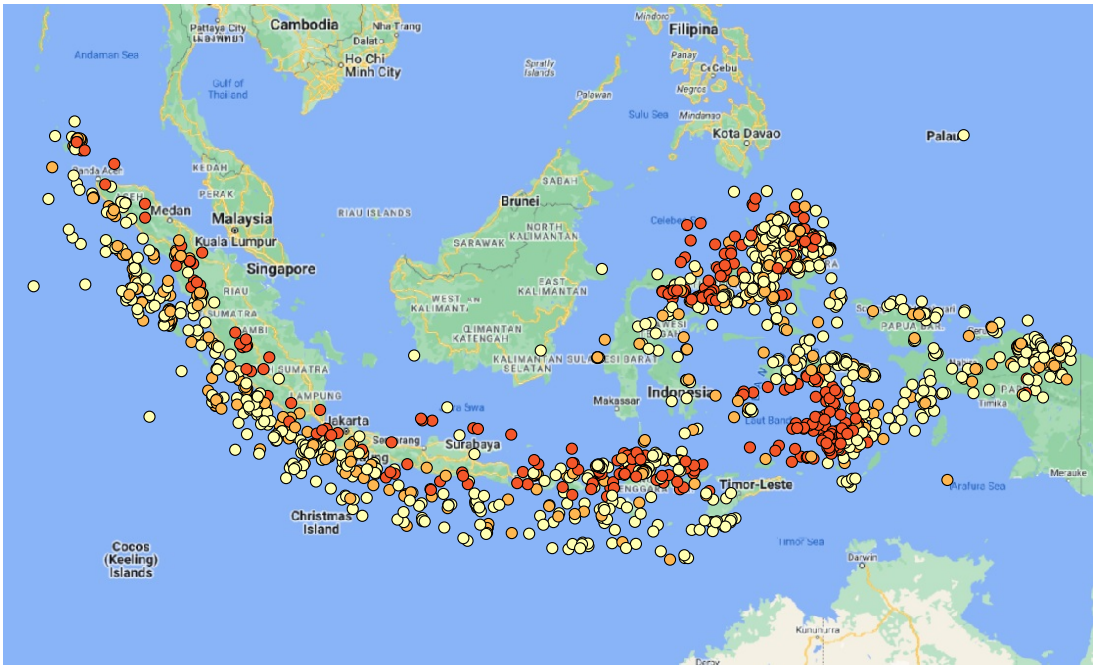


(a) Sabo Model

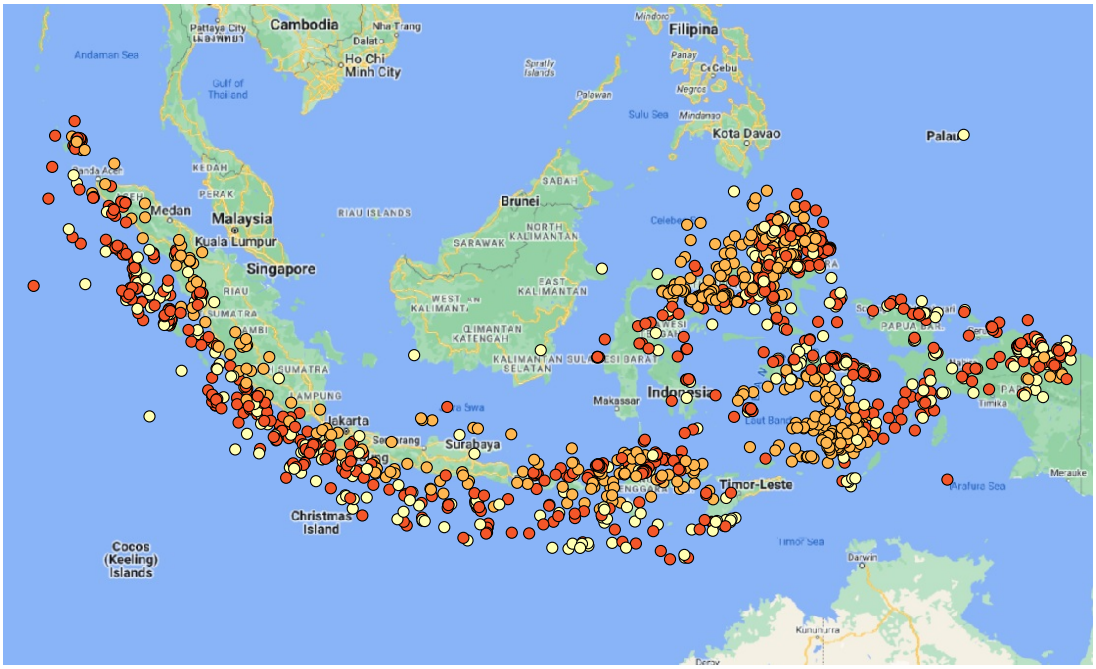


(b) Proposed Model

Figure 2: Clustering results of two clustered-earthquake areas: (1) yellow for cluster 1, (2) orange for cluster 2



(a) Sabo Model



(b) Proposed Model

Figure 3: Clustering results of three clustered-earthquake areas: (1) yellow for cluster 1, (2) orange for cluster 2, and (3) red for cluster 3

Table 4: Clustering profile

2 Clusters				
Model	Cluster	Number of members	Depth average	Magnitude average
Sabo	1	1205 areas	31.394	4.560
	2	481 areas	188.169	4.353
Proposed	1	784 areas	59.786	4.775
	2	902 areas	90.319	4.262
3 Clusters				
Model	Cluster	Number of members	Depth average	Magnitude average
Sabo	1	869 areas	30.386	4.367
	2	376 areas	44.291	5.009
	3	441 areas	193.380	4.331
Proposed	1	336 areas	24.082	4.227
	2	624 areas	162.696	4.370
	3	726 areas	25.793	4.740

The clustering profile results from the Sabo method show a more uneven distribution in the number of members per cluster compared to the proposed method. Additionally, the significant differences in the average depth and magnitude of earthquakes between clusters indicate that the Sabo method's clustering differentiates earthquake-prone areas less specifically than the proposed method. This suggests that the proposed method provides a more refined and accurate classification of the various earthquake zones based on the depth and magnitude characteristics.

6 Conclusion

In this article, a new general form of global optimization model with a non-exponential and non-logarithmic functions has been formulated. The proposed model aims to solve the data clustering problem. The suggested general form enhances the versatility of the model. Our proposed model can overcome some of the drawbacks of previous optimization models. The derivation of its mathematical properties also shows that our proposed function is a closed and bounded function so that we can guarantee that its global minimum exists. Based on numerical simulations, we have found that our proposed model is computationally efficient, measured by the number of evaluation functions and the number of optimal iterations. Furthermore, application to clustering the earthquake affected areas has been done where the results have also been compared to other competitive models. The results of the applications and comparison show that the proposed model is superior.Other than for clustering the earthquake affected areas, our proposed model can also be applied to other problems with the dimension of more than one.

Acknowledgement

This research is supported by the Ministry of Education, Culture, Research and Technology, Indonesia, through the fundamental research grant 2024 (Grant Number 0459/E5/PG.02.00/2024).

References

Alok, A.K., Gupta, P., Saha, S. & Sharma, V. (2020). Simultaneous feature selection and clustering of micro-array and rna-sequence gene expression data using multiobjective optimization. *International Journal of Machine Learning and Cybernetics*, *11*, 2541–2563.

Bandeen-Roche, K., Miglioretti, D.L., Zeger, S.L. & Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, *92*, 1375–1386.

Bjorkman, M., Holmstrom, K. (1999). Global optimization using the direct algorithm in Matlab. *Advanced Modeling and Optimization*, *1*, 17–37.

Caroline, X.G., Dominic, D.Z.Y., Catherine, L.S., Lan, D., Kate, M.F., Johanna, B., Jana, M.M., Teresa, W., Christoph, B., Stephen, W. & Sue, M.C. (2023). An overview of clustering methods with guidelines for application in mental health researc. *Psychiatry Research*, *327*, 115265.

Costa, M.F.P., Rocha, A.M.A.C. & Fernandes, E.M.G.P. (2018). Filter-based direct method for constrained global optimization. *Journal of Global Optimization* , *71*, 517–536.

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D.L. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*, 1–18.

D'Urso, P., Disegna, M., Massari, R., & Prayag, G. (2015). Bagged fuzzy clustering for fuzzy data: An application to a tourism market. *Knowledge-Based Systems*, *73*, 335–346.

D'Urso, P., Giovanni, L.D., Massari, R., Decclesia, R.L. & Maharaj, E.A. (2020). Cepstral-based clustering of financial time series. *Expert Systems with Applications*, *161*, 113705.

France, S.L., Ghose, S. (2019). Marketing analytics: Methods, practice, implementation, and links to other fields. *Expert Systems with Applications*, *119*, 456–475.

Furno, M. (2023). Computing finite mixture estimators in the tails. *Journal of Classification*, *40*, 267–297.

Gablonsky, J.M., Kelley, C.T. (2001). A locally-biased form of the direct algorithm. *Journal of Global Optimization* , *21*, 27–37.

Gao, C.X., Dwyer, D., Zhu, Y., Smith, C.L., Du, L., Filia, K.M., Bayer, J., Menssink, J.M., Wang, T., Bergmeir, C., Wood, S. & Cotton, S.M. (2023). An overview of clustering methods with guidelines for application in mental health research. *Psychiatry Research*, *327*, 115265.

Guha, S., Rastogi, R. & Shim, K. (2000). Rock: A robust clustering algorithm for categorical attributes. *Information Systems*, *25*, 345–366.

Guo, R., Chen, J. & Wang, L. (2021). Hierarchical k-means clustering for registration of multi-view point sets. *Computers & Electrical Engineering*, *94*, 107321.

Guan, X., Terada, Y. (2023). Sparse kernel k-means for high-dimensional data. *Pattern Recognition*, *144*, 109873.

Hashemi, S.E., Jouybari, F.G. & Keshteli, M.H. (2023). A fuzzy c-means algorithm for optimizing data clustering. *Expert Systems with Applications*, *227*, 120377.

Hodges, K., Wotring, J. (2000). Client typology based on functioning across domains using the cafas: Implications for service planning. *The Journal of Behavioral Health Services & Research*, *27*, 257–270.

Iyigun, C., Ben-Israel, A. (2010). A generalized weiszfeld method for the multi-facility location problem. *Operations Research Letters*, *38*, 207–214.

Jain, A.K., Murty, M.N. & Flynn, P.J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*, 264–323.

Jain, A.K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, *31*, 652–666

Jones, D.R., Perttunen, C.D. & Stuckman, B.E. (1993). Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, *79*, 157–181.

Kavitha, E., Tamilarasan, R. (2020). Agglo-hi clustering algorithm for gene expression micro array data using proximity measures. *Multimedia Tools and Applications*, *79*, 1573–7721.

Latifi-Pakdehi, A., Daneshpour, N. (2021). Dbhc: A dbscan-based hierarchical clustering algorithm. *Data & Knowledge Engineering*, *135*, 101922.

Liang, Z., Chen, P. (2021). An automatic clustering algorithm based on the density-peak framework and chameleon method. *Pattern Recognition Letters*, *150*, 40–48.

Liu, H., Wang, Y., Guan, S. & Liu, X. (2017). A new filled function method for unconstrained global optimization. *International Journal of Computer Mathematics*, *94*, 2283–2296.

Lovison, A., Miettinen, K. (2021). On the extension of the direct algorithm to multiple objectives. *Journal of Global Optimization*, *79*, 387–412.

Monti, S., Tamayo, P., Mesirov, J. & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*, 91–118.

Maulik, U., Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern Recognition*, *33*, 1455–1465.

Mittal, H., Pandey, A. C., Saraswat, M., Kumar, S., Pal, R. & Modwel, G. (2022). A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimedia Tools and Applications* , *81*, 35001–35026.

Mullner, D. (2013). Fastcluster: fast hierarchical, agglomerative clustering routines for r and python. *Journal of Statistical Software*, *53*, 1–18.

Nasiri, J., Khiyabani, F.M. (2018). A whale optimization algorithm (woa) approach for clustering. *Cogent Mathematics & Statistics* , *5*, 1483565.

Oyewole, G.J., Thopil, G.A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, *56*, 6439–6475.

Rahnema, N., Gharehchopogh, F.S. (2020). An improved artificial bee colony algorithm based on whale optimization algorithm for data clustering. *Multimedia Tools and Applications* , *79*, 32169–32194.

Ray, S.S., Bandyopadhyay, S. & Pal, S.K. (2007). Gene ordering in partitive clustering using microarray expressions. *Journal of Biosciences* , *32*, 1019–1025.

Rengasamy, S., Murugesan, S. (2021). Pso based data clustering with a different perception. *Swarm and Evolutionary Computation*, *64*, 100895.

Ricotta, C. (2019). Can we trust the chord (and the hellinger) distance. *Community Ecology*, *20*, 104–106.

Rui, X., Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*, 645–678.

Sabo, K., Scitovski, R. & Vazler, I. (2013). One-dimensional center-based l-clustering method. *Optimization Letters*, *7*, 5–22.

Sharma, R., Ravinder, M. (2023). Remote sensing image segmentation using feature based fusion on fcm clustering algorithm. *Complex & Intelligent Systems*, *9*, 7423–7437.

Scitovski, R., Scitovski, S. (2013). A fast partitioning algorithm and its application to earthquake investigation. *Computers & Geosciences*, *59*, 124–131.

Scitovski, R. (2017). A new global optimization method for a symmetric lipschitz continuous function and the application to searching for a globally optimal partition of a one-dimensional set. *Journal of Global Optimization*, *68*, 713–727.

Schubert, E., Rousseeuw, P.J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the pam, clara, and clarans algorithms. *Information Systems*, *101*, 101804.

Seydoux, L., Balestriero, R., Poli, P., 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.

Selim, S.J., Alsultan, K. (1991). A simulated annealing algorithm for the clustering problem. *Pattern Recognition*, *24*, 1003–1008.

Tang, L., Xu, P., Xue, L., Liu, Y., Yan, M., Chen, A., Hu, S. & Wen, L. (2023). A novel self-attention model based on cosine self-similarity for cancer classification of protein mass spectrometry. *International Journal of Mass Spectrometry*, *494*, 117131.

Teboulle, M. (2007). A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, *8*, 65–102.

Vijay, R.K., Nanda, S.J. (2023). Earthquake pattern analysis using subsequence time series clustering. *Pattern Analysis and Applications*, *26*, 19–37.

Weber, C.M., Ray, D., Valverde, A.A., Clark, J.A. & Sharma, K.S. (2022). Gaussian mixture model clustering algorithms for the analysis of high-precision mass measurements. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated*, *1027*, 166299.

Wang, L., Wang, H., Han, X. & Zhou, W. (2021). A novel adaptive density-based spatial clustering of application with noise based on bird swarm optimization algorithm. *Computer Communications*, *174*, 205–214.

Wu, C., Chen, Y., Dong, Y., Zhou, F., Zhao, Y. & Liang, C. J. (2023). Vizoptics: Getting insights into optics via interactive visual analysis. *Computers and Electrical Engineering*, *107*, 108624.

Xie, W.B., Liu, Z., Das, D., Chen, B. & Srivastava, J. (2023). Scalable clustering by aggregating representatives in hierarchical groups. *Pattern Recognition*, *136*, 109230.

Yamagishi, Y., Saito, K., Hirahara, K. & Ueda, N. (2021). Spatio-temporal clustering of earthquakes based on distribution of magnitudes. *Applied Network Science*, *6*, 71.

Yuan, X., Yu, H., Liang, J. & Xu, B. (2021). A novel density peaks clustering algorithm based on k nearest neighbors with adaptive merging strategy. *International Journal of Machine Learning and Cybernetics*, *12*, 2825–2841.

Zhang, Q., Huang, Q., Dong, J., Wang, A., Sun, J., Lan, Q. & Hu, G. (2005). Gene expression profiling of ganglioglioma malignant progression by cdna array. *Chinese Journal of Cancer Research*, *17*, 11–16.

Zhu, X., Hunter, D.R. (2019). Clustering via finite nonparametric ica mixture models. *Advances in Data Analysis and Classification*, *13*, 65–87.

Zhang, T., Ramakrishnan, R. & Livny, M. (1997). Birch: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, *1*, 141–182.