BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0413–0426
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0413-0426

ANALYZING THE EFFECT OF SIMILARITY FUNCTIONS ON PARTITIONING AROUND MEDOIDS ALGORITHM FOR MAPPING DHF DISEASE IN NORTH SUMATRA

Wahyu Nur Fadillah [1], Yulita Molliq Rangkuti [2*], Ichwanul Muslim Karo Karo [3]

[1,2,3] *Computer Science, Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan*
Willem Iskandar Street, Pasar V, Medan Estate, Medan, 20221, North Sumatera, Indonesia

*Corresponding author's e-mail: * **yulitamolliq@unimed.ac.id***

Article History:

Received: 10th September 2023
Revised: 12th December 2023
Accepted: 9th January 2024

Keywords:

DHF;
PAM;
Silhouette Index;
Similarity Function.

ABSTRACT

Dengue hemorrhagic fever (DHF) is an acute febrile illness caused by a virus through the Aedes mosquito. North Sumatra is among the three provinces with the highest incidence and mortality rates in Indonesia. Mapping of DHF cases is very important in efforts to control and prevent the disease. The Partitioning Around Medoid (PAM) algorithm is commonly used to cluster DHF cases. The idea of PAM is a clustering algorithm with a similarity-based approach to grouping objects in one cluster. There are two main focuses in the research: mapping regencies/cities based on dengue case information and analyzing the performance of several similarity functions. The dataset includes variables of incidence rate (IR), case fatality rate (CFR), larva-free rate (ABJ), and population, obtained from the North Sumatra Provincial Health Office and the Central Statistics Agency (BPS). The analysis showed that three clusters were formed in North Sumatra Province. The first cluster includes regencies/cities such as Langkat, Deli Serdang, Karo, Simalungun, Dairi, Samosir, Humbahas, North Labuhan Batu, North Padang Lawas, South Labuhan Batu, Padang Sidempuan, Nias, South Nias, North Nias, and Sibolga. The second cluster consists of regencies/cities such as Medan, Binjai, Sedang Berdagai, Tebing Tinggi, Batubara, Asahan, Tanjung Balai, Labuhan Batu, Toba, North Tapanuli, Central Tapanuli, Gunungsitoli, and West Nias. The third cluster includes the regencies of South Tapanuli and Mandailing Natal. In addition, an evaluation was conducted using the Silhouette Index to measure the quality of the clustering. Based on the comparison using distance methods (Euclidean distance, Manhattan distance, Minkowski distance, and Chebyshev distance), the highest Silhouette Index value was obtained using Chebyshev distance, which amounted to 0.527554. This value indicates reasonable cluster quality. Thus, this study contributes to the mapping of DHF cases in North Sumatra Province and can be the basis for decision-making in overcoming the disease.

How to cite this article:

W. N. Fadillah, Y. M. Rangkuti and I. M. K. Karo., "ANALYZING THE EFFECT OF SIMILARITY FUNCTIONS ON PARTITIONING AROUND MEDOIDS ALGORITHM FOR MAPPING DHF DISEASE IN NORTH SUMATRA," *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0413-0426, March, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*
Research Article · Open Access

1. INTRODUCTION

Dengue hemorrhagic fever (DHF) is an acute febrile illness caused by the dengue virus [1]. DHF is one of the endemic diseases with many fatalities in Indonesia. DHF is still a public health problem whose number of sufferers continues to increase from year to year. Currently, DHF has been found in all regions of Indonesia. North Sumatra is among the three provinces with the highest incidence and mortality rates of DHF in Indonesia. In the last three years, cases of dengue fever have increased by almost 80% in 2020 from the previous year, and cases have decreased in 2021 to 60%, then experienced an increase of 64% in 2022, with the number of dengue cases reaching 4,856. The rise in dengue fever cases in North Sumatra in 2020 and 2022 is why this research must be carried out. Environmental factors influence the problem of dengue fever. So, cases of dengue fever are closely related to geography or region; by knowing the location of the sufferer's living conditions, one can find out the cause of the dengue fever problem in that area (Ainnurriza et al., 2020). The Health Service in North Sumatra Province does not have a geographic system to monitor areas infected with dengue fever, so research must be conducted to build a dengue fever monitoring system in North Sumatra using a Geographic Information System (GIS). In addition, the process of mapping areas where there are cases of dengue fever in North Sumatra Province is still carried out manually, so the process for analyzing the condition of an area is only seen from the number of residents affected by dengue fever and is not followed by mapping areas that are dangerous or prone to dengue fever. From this condition, a method is needed that can carry out this mapping using the clustering method.

Clustering is one of the tasks mining that can be used for mapping based on similar characteristics [2] [2]**.** Partitioning Around Medoid (PAM) is a well-known clustering algorithm for mapping various cases. The idea of PAM algorithm is to use partition clustering method to group a set of *n* objects into *k* clusters based on their similarity [3]. The number of clusters (*k*) is initialized by the user. Generally, the similarity function in PAM algorithm is Euclidean distance.

Each regency and city in North Sumatra have information on the incidence rate (IR), case fatality rate (CFR), larva-free rate (ABJ) of DHF, and population. The mitigation and treatment of dengue among regions cannot be equalized for all regions due to differences in geographical conditions and data information. So, it is necessary to map the area based on geographical information and DHF data information. The output of the mapping is the formation of regencies and cities into several clusters. The members in one cluster must have similar characteristics. Thus, the mitigation and treatment of DHF are more in line with regional characteristics and more tangible.

Several studies have applied the PAM algorithm to analyze DHF disease. Research by [4] grouped the regions based on the high and low dengue case information in Karawang Regency. The results showed that the PAM algorithm produced a strong cluster with a silhouette index value of 0.78793. Other research focused on clustering DHF-prone areas in Boyolali, Indonesia [5]. This system clusters 3 level zones of endemic and 3 level zones of sporadic based on geographic information systems. The three groups are identified as a strong cluster, with the average coefficient for level 1 being 0.837, level 2 being 0.858, and level 3 being 0.773.

The PAM algorithm has some reliability in the mapping process. In some cases, the PAM algorithm produces better quality clusters compared to other algorithms, such as in the study of [4], [3], [6], and [7]. The algorithm is not affected by the order of the dataset. This algorithm is sensitive to noise and outliers, where objects with large values are likely to deviate from the data distribution [8]. This algorithm can also be implemented on time series dataset clustering. [9]. In addition, the partitioning method has better clustering results than other approaches (density or hierarchy) to cluster data [10].

The similarity function is a critical part of the PAM algorithm. Some researchers have analyzed several similarity functions of PAM. A study analyzed Euclidean distance, Manhattan distance, and Chebyshev distance as similarity functions on PAM for clustering plantation commodities [11]. The best number of clusters produced is *k*=5. Manhattan distance on PAM produces the strong cluster, with a Silhouette Coefficient value of 0.9547019. In another case, similarity function analysis on PAM was also provided by [12]. They analyzed the clustering performance of Euclidean and Gower distance on PAM. Three performance metrics were used: Silhouette, Dunn, and connectivity values. The results showed that Gower distance on PAM is not better than the existing PAM with Euclidean distance.

Based on the conditions described above, this study maps DHF cases in North Sumatra using the PAM algorithm. The dataset variables used include incidence rate (IR), case fatality rate (CFR), larva-free rate (ABJ), and population. In addition, this study also analyzes similarity functions in the PAM algorithm.

The similarity functions are Chebyshev Distance, Manhattan, Euclidean¸ and Minkowski. The clustering results from the combination of PAM and similarity functions are evaluated using the silhouette index.

2. RESEARCH METHODS

Several previous studies have analyzed the effect of similarity functions in clustering algorithms for several cases. A study analyzed Euclidean distance, Manhattan distance, and Chebyshev distance as similarity functions on PAM for clustering plantation commodities [11]. Manhattan distance in the PAM algorithm produces the strong cluster, with silhouette index of 0.955. Another study by [13] analyzed several distance functions in the DBSCAN algorithm, such as Euclidean, L1, Hausdorff, Fréchet, Dynamic Time Warping (DTW), Longest Common SubSequence (LCSS), Edit Distance on Real signals (EDR), and Edit distance with Real Penalty (ERP), in case of trajectory clustering. The results revealed that Euclidean distance has the highest purity index, which proves to show superiority over the other distances. However, EDR distance produces the best quality of clusters. Research by [14] analyzed the effects of various similarity functions on the PAM clustering algorithm in document clustering cases. Their findings conclude that Chi-Square works best for document collection, with an efficiency of around 80%, followed by Canberra and Euclidean distances at 70%. The results also show that distance metrics such as Bray-Curtis, Variational, and Trigonometric functions did not produce good results. Based on the information from previous studies, it is possible to analyze various similarity functions in the PAM algorithm for mapping DHF diseases in North Sumatra.

This section describes the structure of the research. **Figure 1** presents a procedure for the implementation of PAM. There are three important processes to produce the best mapping of DHF in North Sumatra: data pre-processing, clustering process using PAM, and evaluation. Based on the figure, there are four distances such as Euclidean, Manhattan, Minkowski, and Chebyshev distances. Validation of PAM using Silhouette index.

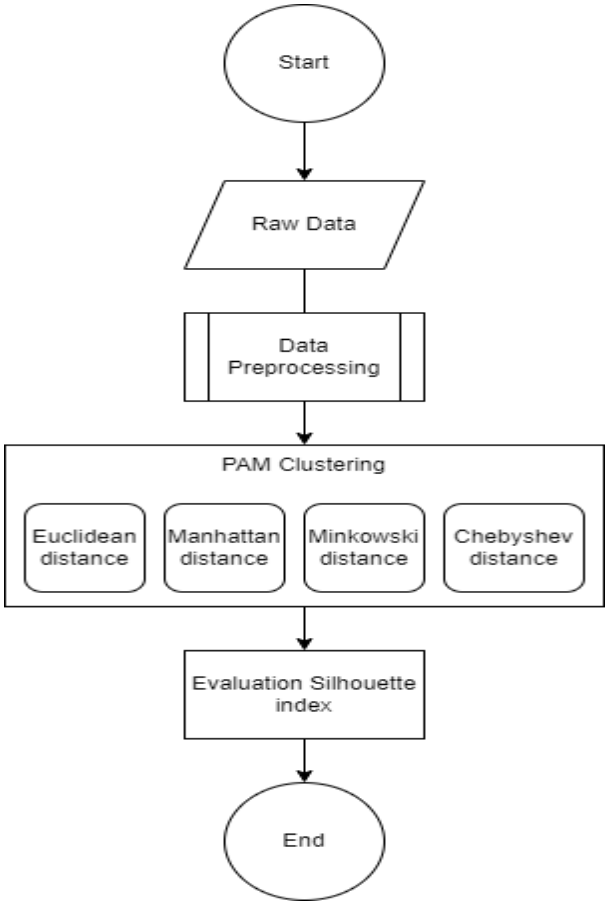

Figure 1. Research Flowchart

2.1 Row Data

North Sumatra is the third provinces with the highest incidence and mortality rates in Indonesia [7]. The data in Figure 2 show that in the last five years, the highest number of DHF cases occurred in 2019, after which it declined and increased again in 2022. The increasing pattern of DHF cases in 2022 was similar to that in 2019. These data prove that DHF is still very dangerous and endemic to public health in North Sumatra.

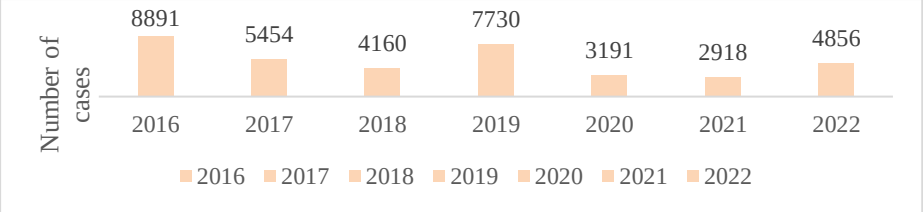

Figure 2. DHF Graph in North Sumatra

The data is also supported by the statement of a North Sumatra Provincial Health Office staff member during an interview session. She stated that "recently, DHF has been in the spotlight and has become very worrisome." In addition, she explained that the mapping of areas based on dengue cases in North Sumatra Province is still conducted manually, so the process of analyzing the condition of an area is only seen from the number of people affected by dengue disease and is not followed by grouping areas that are dangerous or prone to dengue disease. For this reason, a study is needed to map the condition of the spread of dengue disease in each regency/ city in North Sumatra Province. This study used secondary data gathered from the North Sumatra Provincial Health Office and the Central Statistics Agency (BPS) from January 2022 to April 2023. There are four variables collected as raw data: incidence rate (IR), case fatality rate (CFR), and, larva-free rate (ABJ). The dataset is shown in **Table 1**.

Table 1. Row Data

No	City/ Regency	IR	CFR	ABJ	Population
1	Medan	46.0	0.40	95.7	2,460,858
2	Pematang Siantar	98.8	3.97	92.76	270,768
3	Binjai	51.3	-	95.5	295,361
4	Tanjung Balai	44.0	-	94.17	177,640
5	Tebing Tinggi	54.2	0.76	91.76	174,969
6	Sibolga	61.1	-	-	89,932
⋮	⋮	⋮	⋮	⋮	⋮
32	Nias Barat	13.0	1.96	-	90,585
33	Gunungsitoli	183.78	-	-	136,707
Average		30.7	0.69	35.19	

***Data source**: North Sumatra Health Office*

Table 1 presents data on Dengue Hemorrhagic Fever (DHF) cases across 33 districts or cities. These data were obtained from the Provincial Health Office of North Sumatra covering the period from January 2021 to April 2023. It comprises the average incident rate from 2021 to April 2023, the average case fatality rate from 2021 to April 2023, and the average Breteau Index from 2021 to April 2023. However, there were missing data in the CFR and BI columns. This is because the CFR variable is calculated by dividing the number of deaths by the number of cases within a specific timeframe. Therefore, when the number of deaths was zero, the CFR value was left blank. Similarly, the BI variable is derived from outreach activities conducted by the Health Department in specific areas, but these outreach activities are not conducted every month. Consequently, in certain months, the BI value is left blank.

2.2 Data Preprocessing

Data preprocessing is the preparation stage of data in a format suitable for inclusion in the algorithmic process. Data preprocessing is also a critical process in the clustering task. Even in some cases, data preprocessing consumes a lot of energy [15], [16], and [17]. In addition, data preprocessing can improve the quality of clustering research results.

This study applied normalization. Normalization is the process of transforming raw data to have a uniform or balanced scale for each feature. The reasoning is to solve dominant features in dataset distribution and overcome outliers [18]. In this research, the normalization method used is Z-score normalization. The technical normalization process refers to research [19].

2.3 PAM Algorithm

The PAM algorithm was proposed by Kaufman and Rousseeuw [20]. The algorithm is an enhancement of K-Means, and both are in the partition clustering categories. The PAM algorithm is shown in Table 2.

Table 2. PAM Algorithm

PAM Algorithm
Input: D: a dataset that containing n objects number of cluster k
Process: 1) Arbitrarily choose k objects in D as initial medoids. 2) Repeat: 3) Assign each remaining object to the cluster with closest (depend on similarity function result) medoids 4) Randomly, select non-medoids objects, donated as a O_{random}: an object *non-medoid.* 5) Compute total cost, S, of swapping medoids, O_j, O_{random}. 6) If $S < 0$ then swap $O_j$with O_{random} to form new set of k medoids. Until no change
Output: A set of *k* clusters

In the PAM algorithm, determining the value of k is critical. This research applied elbow method with inertia to predict the best number of clusters. Similar things have also been reported in other studies [21]. Figure 3 is the result of predicting the number of clusters using the elbow method. The best number of clusters is determined from the elbow point. From the figure, the elbow point value is 3, which means that the best number of clusters is three. Therefore, the process of analyzing the similarity function on PAM also uses $k = 3$.

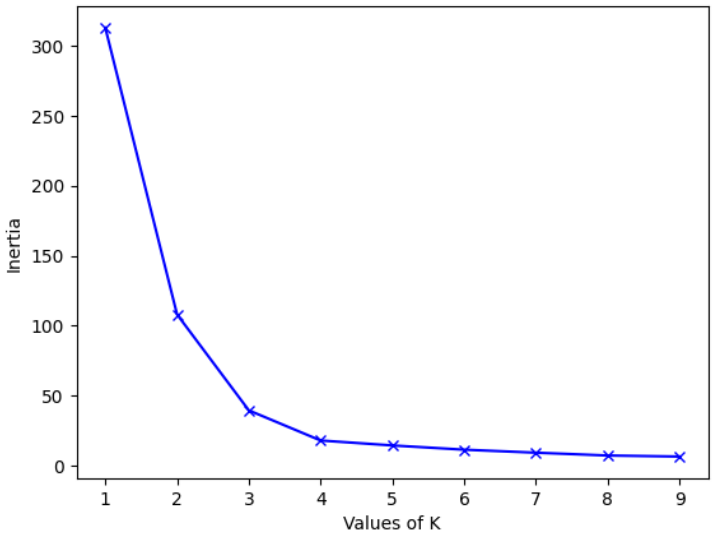

Figure 3. Determining number of k

2.4 Silhouette Index

Silhouette Index (SI) is one of the cluster validation methods through the combination of cohesion and separation values [3], SI was calculated using **Equation (1)**.

$$SI = \frac{1}{n}\sum_{i=1}^{n} s_{(i)} \quad (1)$$

SI is the average silhouette index value for each record ($s_{(i)}$), $a_{(i)}$ is the average dissimilarity of i^{th} object to all other objects in the same cluster (**Equation (4)**), $b_{(i)}$ is the average dissimilarity of i^{th} object with all objects in the closest cluster (**Equation (3)**).

$$s_{(i)} = \frac{b_{(i)} - a_{(i)}}{max\{a_{(i)}, b_{(i)}\}} \quad (2)$$

$$b_{(i)} = \min d_{(i,C)} \quad (3)$$

$$a_{(i)} = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d_{(i,j)} \quad (4)$$

Subjective criteria for judging strong, weak, or not-found clusters by SI are previous research criteria [22]. The criteria are shown in **Table 3**.

Table 3. Silhouette Index Criteria

Silhouette Index (SI)	Criteria
$0.7 < SI \leq 1$	Strong cluster
$0.5 < SI \leq 0.7$	Reasonable cluster
$0.25 < SI \leq 0.5$	Weak cluster
$SI \leq 0.25$	Not found

Table 3 above outlines the Silhouette Index (SI) criteria for assessing the quality of clustering results. The Silhouette Index is a metric used to evaluate the coherence and separation of clusters generated by clustering algorithms. It quantifies the extent to which data points within the same cluster are similar to each other compared to data points in other clusters.

1. Strong cluster ($0.7 < SI \leq 1$): This category indicates that the clusters formed are well-defined and distinct. A Silhouette Index in this range signifies a high degree of separation between clusters and suggests that the data points are correctly assigned to their respective clusters. In such cases, the clustering algorithm has successfully identified meaningful patterns in the data.
2. Reasonable cluster ($0.5 < SI \leq 0.7$): Silhouette Index values falling within this range indicate that the clusters are reasonably well-formed. While there is some overlap or proximity between clusters, the clustering algorithm has managed to create clusters with a moderate level of separation. This may suggest that the clustering results are useful, but there could be room for improvement in terms of cluster distinction.
3. Weak cluster ($0.25 < SI \leq 0.5$): Weak cluster status is associated with Silhouette Index values in this range. It implies that the clustering results exhibit relatively poor separation and might not accurately represent underlying patterns in the data. In such cases, clusters may be less coherent, and improvements in the clustering process may be required.
4. Not found ($SI \leq 0.25$): A Silhouette Index below or equal to 0.25 signifies that the clustering results are not meaningful or that the data might not naturally form distinct clusters. This may indicate that the clustering algorithm struggled to group data points effectively, and further analysis or a different approach may be necessary to uncover meaningful structures in the data.

In summary, the Silhouette Index provides a quantitative means of evaluating clustering quality, with higher values indicating better-defined clusters. Researchers and data analysts can use these criteria to assess the strength of clustering results and make informed decisions about the suitability of the applied clustering algorithm for their specific dataset and research goals.

2.5 Similarity Function

Euclidean distance is a standard similarity function used in various algorithms. The classical PAM algorithm used Euclidean distance as a similarity function, according to the formula shown in **Equation (5)**.

$$d_2(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{5}$$

where $d_2(x,y)$ is the Euclidean distance between object x_i and y_i, x_i is the data in the ith attribute and y is the ith data center, and n represents the number of dimensions **[23]**.

Manhattan distance is a method of calculating the distance between two points based on the sum of the results of the absolute value of the difference in the value of each dimension between the two points **[23]**. **Equation (6)** is to calculate the Manhattan distance using the following formula.

$$d_1(x,y) = \sum_{i=1}^{n}|x_i - y_i|, \tag{6}$$

where $d_1(x,y)$ is the Manhattan distance.

Minkowski distance is generalized Euclidean and Manhattan distance **[24]**. The formula of Minkowski distance is shown in **Equation (7)**. The p parameter represents the order of the norm; it must be an integer **[23]**, There are similarities between the Minkowski distance and Euclidean distance or Manhattan distance. If the order (p) is 1, it will represent Manhattan Distance, and if the p parameter is 2, it will mean Euclidean distance.

$$d_p(x,y) = \sqrt[p]{\sum_{i=1}^{n}|x_i - y_i|^p}, \tag{7}$$

where $d_p(x,y)$ is the Minkowski distance.

Chebyshev distance is a method of measuring the distance between two points based on the absolute value or absolute value of the coordinate difference between the two points. Chebyshev distance can be considered as the maximum distance between two points, which is the distance between two points measured based on the maximum value difference between the x and y coordinates of the two points. **Equation (8)** is the formula to calculate the Chebyshev distance between two objects.

$$d_\infty(x,y) = \max_{i=1}|x_i - y_i|, \tag{8}$$

where $d_\infty(x,y)$ is the Chebyshev distance

3. RESULTS AND DISCUSSION

3.1 Correlation Analysis

Initially, this study presents a correlation analysis between variables. The procedure uses Pearson's correlation. The correlation matrix obtained is shown in **Figure 4**. The positive correlation value is unidirectional; it indicates that when the independent variable is large, it is followed by the dependent variable. While the negative correlation value is the opposite, the independent variable becomes smaller as the dependent variable becomes larger. The correlation value ranges from 0 to 1. If it is close to 1 or -1, it indicates that the relationship between the two variables is getting stronger, and if the correlation value is close to 0, the relationship between the two variables is getting weaker.

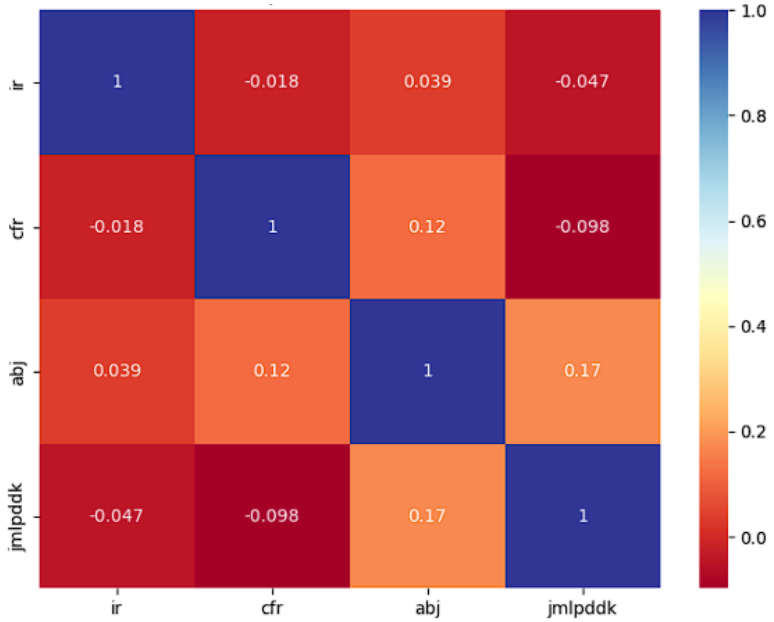

Figure 4. Heatmap Correlation of Variable DHF

Generally, the correlation value of variables in this study is below 0.25. Based on the standards used in previous studies, the correlation between variables is very weak. It means that one variable with another variable has little connection or influence. Thus, all variables are independent variables. The normalized data was used using Z-score.

3.2 Correlation Analysis

Initially, this study presents the classical PAM using Euclidean distance as the similarity function to group the objects. Based on the information of the elbow method, the best number of clusters used is $k = 3$. Therefore, Figure 5 shows the clustering results using classical PAM.

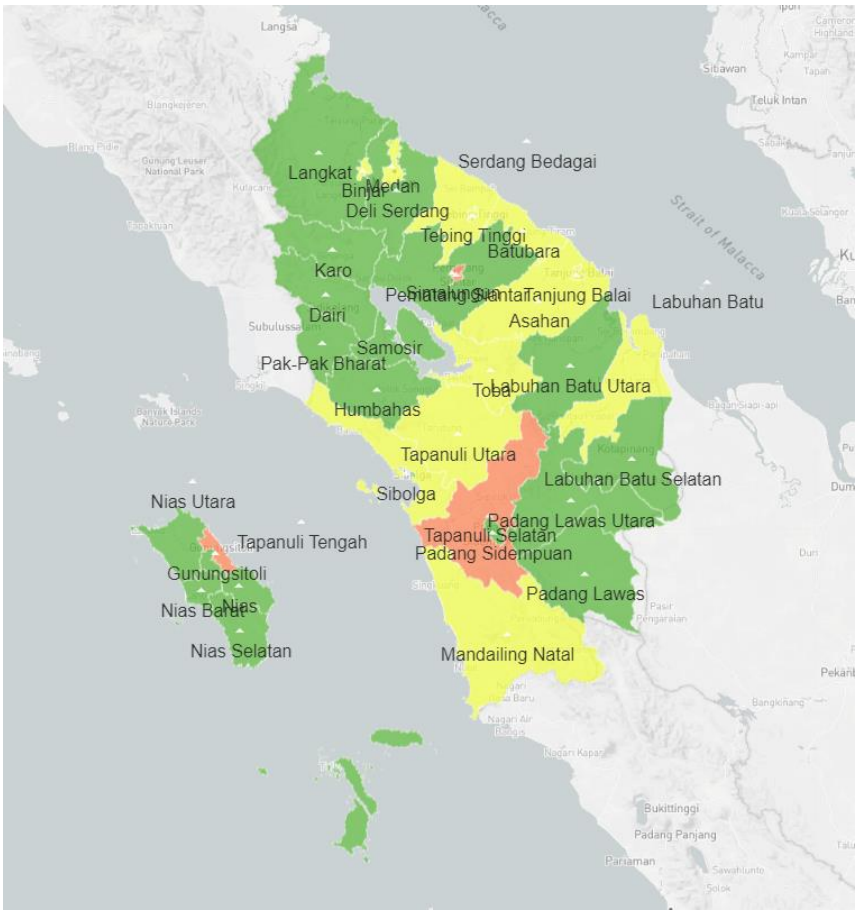

Figure 5. DHF Mapping using Classical PAM Clustering

The color indicates the level of distribution of DHF cases in each area. Green color (1) represents areas with low levels of dengue cases, yellow color (3) is medium level, and orange color (2) is high level. The membership of each cluster can be seen in the classical PAM column table. The quality of the resulting cluster is $SI = 0.340201$. Based on the SI value, the mapping results from classical PAM clustering have a weak cluster.

Table 4. Cluster Membership of each Experiment

City/Regency	Classical PAM	Manhattan +PAM	Minkowski + PAM	Chebyshev + PAM
Medan	3	3	3	3
Pematang Siantar	2	2	2	2
Binjai	3	3	3	3
Tanjung Balai	3	3	3	3
Tebing Tinggi	3	3	3	3
Sibolga	1	1	1	1
Padang Sidempuan	1	1	1	1
Deli Serdang	1	1	1	1
Langkat	1	1	1	1
Karo	1	1	1	1
Simalungun	1	1	1	1
Asahan	3	3	3	3
Labuhan Batu	3	3	3	3
Tapanuli Utara	3	3	3	3

City/Regency	Classical PAM	Manhattan +PAM	Minkowski + PAM	Chebyshev + PAM
Tapanuli Tengah	3	3	3	3
Tapanuli Selatan	**2**	**1**	**2**	**2**
Nias	1	1	1	1
Dairi	1	1	1	1
Toba	**3**	**3**	**3**	**2**
Mandailing Natal	**3**	**3**	**3**	**1**
Nias Selatan	1	1	1	1
Pak-Pak Bharat	1	1	1	1
Humbahas	1	1	1	1
Samosir	1	1	1	1
Serdang Bedagai	3	3	3	3
Batubara	3	3	3	3
Padang Lawas	1	1	1	1
Padang Lawas Utara	1	1	1	1
Labuhan Batu Selatan	1	1	1	1
Labuhan Batu Utara	1	1	1	1
Nias Utara	1	1	1	1
Nias Barat	1	1	1	1
Gunungsitoli	**2**	**1**	**2**	**2**

3.3 Mapping Results using Manhattan Distance with PAM

In this experiment, the Euclidean distance was replaced by the Manhattan distance as the similarity function for the PAM algorithm. The number of clusters used is still the same as in the previous experiment. The visualization of the mapping can be seen in **Figure 6**, and the cluster membership results can be seen in **Table 4**, column Manhattan + PAM. The mapping result with the PAM algorithm and Manhattan distance has $SI = 0.360119$. The color information and cluster quality are still the same as in the previous experiment. This experiment still produces weak clusters. There are two regions that have changed their membership compared to the previous results. South Tapanulis Regency and Gunung Sitoli City, which were originally included in the orange cluster, are now included in the green cluster.

Figure 6. **DHF Mapping Using Manhattan Distance on PAM Clustering**

3.4 Mapping Results Using Minkowski Distance with PAM

This subsection presents PAM clustering results using Minkowski distance as the similarity function. The number of clusters is still $k = 3$. Visually and in terms of cluster membership, the clustering results of this experiment are the same as the classical PAM, except that the SI value is slightly different, which is 0.340201.

3.5 Mapping Results Using Chebyshev Distance with PAM

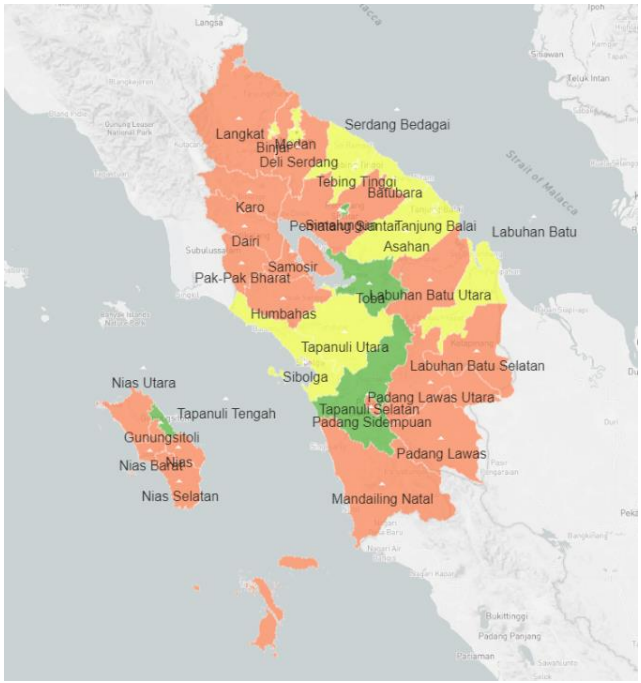

Figure 7. **DHF Mapping Using Chebyshev Distance on PAM Clustering**

In the final experiment, the subsection presents the results of PAM clustering with Chebyshev distance as the similarity function. The number of clusters is still $k = 3$. The mapping visualization of the clustering results can be seen in **Figure 7**, and the cluster membership results can be seen in **Table 4**, column Chebyshev + PAM. There are two city regencies/cities that are different from the mapping results using classical PAM. These are Toba Regency and Mandailing Natal. The membership of Toba Regency changes from cluster yellow to cluster orange. In comparison, the membership of Mandailing changed from a yellow cluster to a green cluster. These changes resulted in a reasonable cluster with $SI = 0.527554$.

3.6 Analysis of Similarity Functions on PAM

This research uses the Silhouette Index to evaluate the clustering results of each experiment. The experiments are a combination of the PAM algorithm and similarity functions. Four similarity functions are analyzed: Euclidean distance, Manhattan distance, Minkowski distance, and Chebyshev distance. The number of clusters studied is $k = 2,3,4,5$ and 6. **Figure 8** presents the performance for each similarity function.

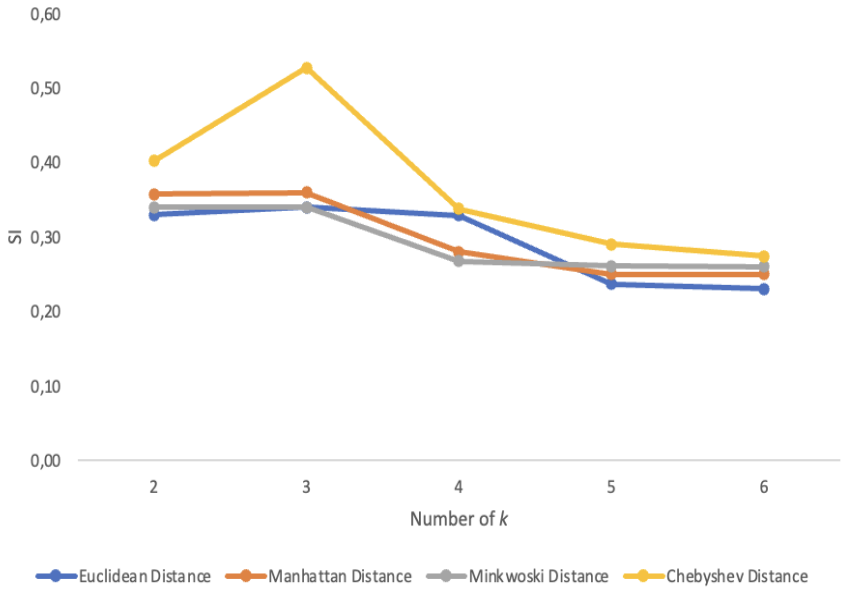

Figure 8. DHF Mapping Using Chebyshev Distance on PAM Clustering

As shown in **Figure 8**, all experiments agree that the best number of clusters is $k = 3$. It can be seen from the SI value of $k = 3$ for each experiment that it is always higher than the other number of clusters. All experiments also agree that the cluster quality matches $k = 3$. Therefore, a large number of k can be tested without being tested. Lastly, the SI value of the Chebyshev Distance on the PAM algorithm is always higher than the other similarity function combinations on PAM in every number of k. Thus, Chebyshev distance as a similarity function on the PAM algorithm is the best for mapping DHF data in North Sumatra.

The implication of all the experiments run is that there are three clusters of DHF mapping areas in North Sumatra. In other words, all combinations of the PAM algorithm and similarity functions recommend the North Sumatra region into three clusters. In fact, the results of the three clusters from each experiment are not the same, the members of each cluster from experiment 3.3 with the cluster members produced by experiments 3.4 and 3.5 are not identical. The PAM algorithm and Chebyshev distance map the North Sumatra region with the best quality.

4. CONCLUSIONS

The PAM algorithm can be implemented to map DHF cases in districts/cities in North Sumatra Province. There were four experiments based on several similarity functions (Euclidean, Manhattan, Minkowski, and Chebychev). All experiments agreed that the best number of clusters was three. From the evaluation results with SI, the Chebyshev Distance is the best distance, so this research uses the Chebyshev

Distance as a distance calculation in the PAM algorithm. This is a new discovery that from research by Rani et al. (2022) the distance that can be recommended is still the Euclidean distance. This number became the reference for mapping areas based on the resulting cluster coloring. A green cluster is an area with a low level of DHF cases, a yellow cluster is a medium cluster, and an orange cluster is a high cluster. The green cluster consists of Langkat, Deli Serdang, Karo, Simalungun, Dairi, Samosir, Humbahas, North Labuhan Batu, North Padang Lawas, South Labuhan Batu, Padang Sidempuan, Nias, South Nias, North Nias, and Sibolga. The yellow cluster consists of Medan, Binjai, Sedang Berdagai, Tebing Tinggi, Batubara, Asahan, Tanjung Balai, Labuhan Batu, Toba, North Tapanuli, Central Tapanuli, Gunungsitoli, West Nias. The orange clusters include South Tapanuli and Mandailing Natal. Based on the silhouette index, the Chebychev distance on the PAM algorithm produces reasonable clusters, while other similarity functions have weak clusters. Thus, Chebychev distance is the best similarity function on PAM for mapping DHF data in North Sumatra than others. Future research can compare cluster results using various PAM clustering algorithms with the Diana method by testing the best distance among Eulidean, Manhattan, and Chebyshev distances.

REFERENCES

[1] N. Y. Lindawati, L. Murtisiwi, T. A. Rahmania, P. N. Damayanti, and F. M. Widyasari, "UPAYA PENINGKATAN PENGETAHUAN MASYARAKAT DALAM RANGKA PENCEGAHAN DAN PENANGGULANGAN DBD DI DESA DLINGO, MOJOSONGO, BOYOLALI," *SELAPARANG Jurnal Pengabdian Masyarakat Berkemajuan*, vol. 4, no. 2, 2021, doi: 10.31764/jpmb.v4i2.4305.

[2] R. Deni and A. Kurnianto, "Data Mapping System Of Riau Province Fire Potential Using K-Means Clustering Method," *JAIA - Journal of Artificial Intelligence and Applications*, vol. 1, no. 1, 2020, doi: 10.33372/jaia.v1i1.640.

[3] W. N. Fadillah, Y. M. Rangkuti, I. Muslim, and K. Karo, "Implementasi Partitioning Around Medoids Pada Visualisasi Penyebaran Penyakit DBD di Sumatera Utara," *Journal of Mathematics*, vol. 6, no. 2, pp. 128–137, 2023, doi: https://doi.org/10.35580/jmathcos.v6i2.52350.

[4] D. R. Agustian and B. A. Darmawan, "ANALISIS CLUSTERING DEMAM BERDARAH DENGUE DENGAN ALGORITMA K-MEDOIDS (STUDI KASUS KABUPATEN KARAWANG)," *JIKO (Jurnal Informatika dan Komputer)*, vol. 6, no. 1, 2022, doi: 10.26798/jiko.v6i1.504.

[5] S. Suprihatin, Y. R. W. Utami, and D. Nugroho, "K-MEANS CLUSTERING UNTUK PEMETAAN DAERAH RAWAN DEMAM BERDARAH," *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, vol. 7, no. 1, 2019, doi: 10.30646/tikomsin.v7i1.408.

[6] S. Wulandari and N. Dwitiyanti, "Implementasi Algoritma Clustering Partitioning Around Medoid (PAM) dalam Clustering Virus MERS-Cov," *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, vol. 5, no. 1, 2020, doi: 10.30998/string.v5i1.6469.

[7] T. Akbar, G. M. Tinungki, and S. Siswanto, "PERFORMANCE COMPARISON OF K-MEDOIDS AND DENSITY BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE USING SILHOUETTE COEFFICIENT TEST," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 3, pp. 1605–1616, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1605-1616.

[8] S. Wulandari and N. Dwitiyanti, "IMPLEMENTASI ALGORITMA CLUSTERING PARTITIONING AROUND MEDOID (PAM) DALAM CLUSTERING VIRUS MERS-CoV," 2020. [Online]. Available: www.ncbi.nlm.nih.gov.

[9] C. Cindy, C. Cynthia, V. Vito, D. Sarwinda, B. D. Handari, and G. F. Hertono, "Cluster Analysis on Dengue Incidence and Weather Data Using K-Medoids and Fuzzy C-Means Clustering Algorithms (Case Study: Spread of Dengue in the DKI Jakarta Province)," *Journal of Mathematical and Fundamental Sciences*, vol. 53, no. 3, 2022, doi: 10.5614/j.math.fund.sci.2021.53.3.9.

[10] I. M. K. Karo and A. F. Huda, "Spatial clustering for determining rescue shelter of flood disaster in South Bandung using CLARANS Algorithm with Polygon Dissimilarity Function," in *Proceedings - 2016 12th International Conference on Mathematics, Statistics, and Their Applications, ICMSA 2016: In Conjunction with the 6th Annual International Conference of Syiah Kuala University*, 2017. doi: 10.1109/ICMSA.2016.7954311.

[11] D. D. Abdurrahman, F. Agus, and G. M. Putra, "Implementasi Algoritma Partitioning Around Medoids (PAM) untuk Mengelompokkan Hasil Produksi Komoditi Perkebunan (Studi Kasus: Dinas Perkebunan Provinsi Kalimantan Timur)," *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, vol. 16, no. 2, 2021, doi: 10.30872/jim.v16i2.6520.

[12] A. Aditya, B. Nurina Sari, and T. Nur Padilah, "Perbandingan pengukuran jarak Euclidean dan Gower pada klaster k-medoids," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 1, pp. 1–7, 2021, doi: 10.14710/jtsiskom.2021.13747.

[13] A. Moayedi, R. A. Abbaspour, and A. Chehreghan, "An evaluation of the efficiency of similarity functions in density-based clustering of spatial trajectories," *Ann GIS*, vol. 25, no. 4, 2019, doi: 10.1080/19475683.2019.1679254.

[14] K. Taghva and R. Veni, "Effects of similarity metrics on document clustering," in *ITNG2010 - 7th International Conference on Information Technology: New Generations*, 2010. doi: 10.1109/ITNG.2010.65.

[15] I. M. K. Karo, A. F. Huda, and K. MaulanaAdhinugraha, "A cluster validity for spatial clustering based on davies bouldin index and Polygon Dissimilarity function," in *Proceedings of the 2nd International Conference on Informatics and Computing, ICIC 2017*, 2018. doi: 10.1109/IAC.2017.8280572.

[16] L. R. Costella Pessutto, D. Suarez Vargas, and V. P. Moreira, "Clustering Multilingual Aspect Phrases for Sentiment Analysis," in *Proceedings - 2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018*, 2019. doi: 10.1109/WI.2018.00-91.

[17] I. M. Karo Karo, A. Yusmanto, and R. Setiawan, "Segmentasi Nasabah Kartu Kredit Berdasarkan Perilaku Penggunaan Kartu Kreditnya Menggunakan Algoritma K-Means," *Journal of Software Engineering, Information and Communication Technology*, vol. 2, no. 2, pp. 101–107, 2021, [Online]. Available: https://www.kaggle.com/arjunbhasin2013/ccdata.

[18] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl Soft Comput*, vol. 97, 2020, doi: 10.1016/j.asoc.2019.105524.

[19] I. M. Karo Karo and H. Hendriyana, "Klasifikasi Penderita Diabetes menggunakan Algoritma Machine Learning dan Z-Score," *Jurnal Teknologi Terpadu*, vol. 8, no. 2, pp. 94–99, 2022.

[20] L. A. Ibrahim and I. Fekete, "What machine learning can tell us about the role of language dominance in the diagnostic accuracy of German LITMUS non-word and sentence repetition tasks," *Front Psychol*, vol. 9, no. JAN, 2019, doi: 10.3389/fpsyg.2018.02757.

[21] S. A. P. Raj and Vidyaathulasiraman, "Determining Optimal Number of K for e-Learning Groups Clustered using K-Medoid," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120644.

[22] I. M. Karo Karo, S. Dewi, M. Mardiana, F. Ramadhani, and P. Harliana, "K-Means and K-Medoids Algorithm Comparison for Clustering Forest Fire Location in Indonesia," *Jurnal Ecotipe (Electronic, Control, Telecommunication, Information, and Power Engineering)*, vol. 10, no. 1, 2023, doi: 10.33019/jurnalecotipe.v10i1.3896.

[23] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 4, no. 1, 2019, doi: 10.30591/jpit.v4i1.1253.

[24] I. M. K. Karo, A. Khosuri, and R. Setiawan, "Effects of Distance Measurement Methods in K-Nearest Neighbor Algorithm to Select Indonesia Smart Card Recipient," in *2021 International Conference on Data Science and Its Applications, ICoDSA 2021*, 2021. doi: 10.1109/ICoDSA53588.2021.9617476.