Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 9, No. 3, August, Pp. 287-296

Fuzzy C-Means algorithm modification based on distance measurement for river water quality

Shofwatul Uyun'[*1], Eka Sulistiyowati[2], Tirta Agung Jati[1]
Department of Informatics, Universitas Islam Negeri Sunan Kalijaga Yogyakarta, Indonesia[1]
Department of Biology, Universitas Islam Negeri Sunan Kalijaga Yogyakarta, Indonesia[1]

Article Info
Keywords:
Capacity, Pollution Load; Fuzzy C-Means, Partition Coefficient, Partition Entropy, Silhouette Score

Article history:
Received: February 23, 2024
Accepted: June 25, 2024
Published: August 31, 2024

Cite:
S. 'Uyun, Eka Sulistiyowati, and T. A. Jati, "Fuzzy C-Means Algorithm Modification Based on Distance Measurement for River Water Quality", KINETIK, vol. 9, no. 3, Aug. 2024.
https://doi.org/10.22219/kinetik.v9i3.1991

*Corresponding author.
Shofwatu Uyun'
E-mail address:
shofwatul.uyun@uin-suka.ac.id

Abstract
River water quality could be determined by understanding the capacity of pollutants in a water body. Fuzzy C-Means (FCM) is one of the fuzzy clustering methods for determining river water quality by measuring water quality parameters, that is, dissolved oxygen (DO) and total dissolved solids (TDS). The FCM algorithm is an effective fuzzy clustering algorithm for grouping data but often produces local and inconsistent optimal solutions due to the partition matrix's random initialisation process. Therefore, this study proposes to modify the FCM algorithm to be precise in the partition matrix initialisation process using several distance concepts. The purpose of the proposed algorithm modification is to get more consistent FCM clustering results and minimise stop iterations. The validation process for the clustering results uses the FCM algorithm, and the FCM modification algorithm uses three parameters, namely the Partition Coefficient Index (PCI), Partition Entropy Index (PEI) and Silhouette Score (SS). The experiments were conducted with three replications and using various distance concepts. The results showed that the number of iterations stopped in the FCM algorithm has different values for PCI, PEI, SS, and stop iterations and objective functions in each trial. On the contrary, the FCM modification algorithm has consistent PCI, PEI, and SS values, and the number of iterations stops with fewer iterations. Therefore, the modified algorithm for initialising the partition matrix can be used in the fuzzy C-means clustering algorithm.

1. Introduction

As an archipelago, Indonesia has many rivers, including 3,137 nationally recognised rivers, spread from Sabang to Merauke [1]. The Gajahwong River is one of the three main rivers in Yogyakarta City, Indonesia. The river is known for its long history, attributed to the civilisation of the Sultanate of Ngayogyakarta Hadiningrat. According to the Watershed Management Board, Gajahwong is a part of the Opak Watershed. Very close to Gajahwong, there are two other major rivers, the Winongo and the Code. The downstream of these three rivers is the Opak River, so in general, the watershed is named the Opak River [2]. The river remains an important source of drinking water for many rural and urban inhabitants [3]. Water is not only drawn to meet the drinking water supply, but people also rely on it to maintain cleanliness, sanitation, and hygiene [4]. Unfortunately, river water pollution still poses major challenges in developing countries such as Indonesia. Pollutants induce waterborne diseases such as diarrhoea and dysentery. Domestic and household pollution contribute to the incidence of waterborne diseases [5]. As reported by [4], about 7.5% of deaths were related to waterborne diseases, especially diarrhoea. Although the government has taken steps to ensure that clean energy is accessible, including controlling the pollutant loads that enter the water, it remains a great challenge to overcome the problems because water pollution has become widespread and has taken many forms, including diffuse and non-diffuse pollution [6]. In terms of pollution control, the government has a set of measurement guidelines in the form of water quality standards [7].

In Indonesia, water quality standards have been regulated by Government Regulation No. 82/2001. We also have a more detailed regulation regarding the pollution load, namely SK.298/Menlhk/Setjen/PKL.1/2017. Measurement of the pollution load is important to calculate the number of pollutants entering the water body. Mathematically, the pollution load is defined as the mass of pollutants in a given time (the unit is kg/day). Measurement of pollution load is important to control pollution, as well as for the government to issue a permit for discharge of liquid waste into the river [8]. However, the monitoring of the pollution load is rarely carried out by the authority. They are due to many factors, including the limitation of human resources. Furthermore, pollution load data is rarely published and is taken in the form of academic or regulatory context, which is difficult for lay people to access [9]. Therefore, we need a novel approach to determine pollutant loads and present the results in an accessible way. In this research, we proposed the use of

automation in the system to determine pollutant loads and the river capacity of pollutant loads. In general, there are two systems of automatization: supervised and unsupervised learning [10].

Similarly, there are also two approaches to algorithmic clustering, namely hard clustering and fuzzy clustering. In principle, the difference in the performance of the two algorithms is the membership of each data point in the cluster [11], [12]. In hard clustering, for each piece of data, there is only one cluster member with a full membership value [13]. In fuzzy clustering, a subset of data could be a member of more than one cluster with a membership value between 0 and 1 [14], [15]. Fuzzy C-means Clustering is an effective fuzzy clustering algorithm for grouping data, but often produces local optimal solutions [16], [17]. FCM has more flexible and fair advantages in data treatment compared to conventional clustering algorithms or hard clustering [18]. In addition, the benefit of this algorithm is that it is unsupervised and can reach convergent cluster centres [19]. FCM clustering results in fuzzy rules in fuzzy inference systems [20] and deep learning algorithms [21], [22].

Several previous studies have optimised the FCM algorithm. For example, Surono and Putri [23] performed the FCM optimisation by combining two distance concepts, namely Minkowski and Chebychev (FMMC), using principal component analysis (PCA). PCA is used to reduce the data dimensions to help stabilise the cluster analysis measurement results. The evaluation of clustering performance results is measured using the Davies-Bouldin index (DBI) parameter. The parameter shows an increase in performance from the collaborative use of FCMMC and PCA. Meanwhile, [24] combines the Minkowski and Chebychev distance concepts used in the k-Nearest Neighbours (k-NN) classification algorithm. On the basis of the experimental results, the distance concept can increase the efficiency of the k-NN algorithm. In another study, [25] uses the distance concept based on Kalman filtering in the k-means clustering algorithm applied to deep neural networks.

Some of the weaknesses of the FCM algorithm include: sensitivity to cluster centres; sensitive cluster centres make the final results difficult to control; and FCM often produces local optimal solutions [26]. Furthermore, what causes inconsistent clustering results is that the matrix initialisation process is performed randomly at the beginning of the process [27]. Another important thing that is key to the success of clustering results using the FCM algorithm is the mechanism to calculate the distance between each data point and all cluster centres. The Euclidean concept is the most used distance concept in the FCM algorithm. Several studies have modified it, including Minkowski and Chebychev [23], [24], [28], [29] Minkowski metric [21], and Mahalanobis and Minkowski distance metrics [30].

On the basis of the background above, it is necessary to modify the FCM algorithm to determine the capacity of river pollution loads by comparing the performance of several distance concepts. This modification aims to minimise inconsistent clustering results due to the random initialisation of the matrix. We implemented the modified FCM algorithm proposed on several distance concepts, including Euclidean, Manhattan, Minkowski, Chebyshev, squared Euclidean, Canberra and a combination of Minkowski and Chebyshev. The contribution of this paper is to improve the performance of the FCM algorithm to reduce the number of epochs and obtain more optimal clustering results. The proposed model compares its performance with the FCM algorithm in general using three parameters: partition entropy index (PEI), partition coefficient index (PCI), and silhouette score (SS). This paper is structured as follows. Section 2 describes the proposed method, followed by Section 3, which contains the results and discussion of the proposed method, and finally Section 4, which contains the conclusions of this paper.

2. Proposed Method

This study uses water sample data from the Gadjah Wong River in Yogyakarta to determine the carrying capacity of the pollution load of the river. This study analysed the ability of the load carrying capacity of the pollution with two parameters: dissolved oxygen (DO) and total dissolved solids (TDS). Determining the carrying capacity of river pollution loads for these two parameters uses the mass balance method. The clustering process uses the resulting data from the measurements of each parameter. This study proposes a technique consisting of three stages, namely pre-processing, clustering using the FCM algorithm and modification of the FCM algorithm, and evaluation of the clustering performance of the two algorithms using three parameters, namely: partition entropy index (PEI), partition coefficient index (PCI), and silhouette score (SS).

2.1 Pre-processing

There are 24 sampling locations for the Gadjah Wong river water data, divided into five segments. Measurement of the water discharge and concentration constituents of DO and TDS at each point, while the aim is to measure the cross-sectional area and velocity of the container to determine the water discharge at each end. Therefore, the clustering process for DO and TDS each consists of 5 parameters, including water discharge, DO or TDS concentration in the stream, river pollutant load (kg/day), river pollutant load according to quality standards (kg/day) and capacity allocation. The data for DO is shown in Table 1. For example, point 1, which is part of segment one in the Santo Thomas area, has a water debit of 0.92; the concentration of DO in the stream is 4.8 with a river pollutant load of 414.72 kg/day; and the river pollutant load according to the quality standard of 398.33 kg/day has a capacity allocation of -16.39 based on the mass balance. The TDS values are presented in Table 2.

Table 1. Example of Measurement Result Data for Dissolved Oxygen (DO)

Point	The parameters				
	Water discharge	DO concentration in the flow	Pollutant load in the river (kg/day)	River pollutant load according to quality standards kg/day	Allocation of capacity
1	0.92	4.8	414.72	398.33	-16.39
2	0.79	5.3	457.92	339.25	-118.67
3	1.07	5.8	501.12	463.11	-38.01
4	1.60	5.6	483.84	692.33	208.49
5	0.94	5.1	440.64	404.39	-36.25
6	1.03	4.7	406.08	444.28	38.20
7	0.88	4.6	397.44	380.22	-17.22
8	1.15	5.7	492.48	496.35	3.87

Table 2. Example of Measurement Result Data for Total Dissolved Solid (TDS)

Point	The parameters				
	Water discharge	TDS concentration in the flow	Pollutant load in the river (kg / day)	River pollutant load according to quality standards kg/day	Allocation of capacity
1	0.92	205	17712	79665.38	61953.38
2	0.79	204	17625.6	67850.21	50224.61
3	1.07	228	19699.2	92622.30	72923.10
4	1.60	204	17625.6	138465.90	120840.30
5	0.94	207	17884.8	80877.21	62992.41
6	1.03	206	17798.4	88856.71	71058.31
7	0.88	203	17539.2	76044.09	58504.89
8	1.15	201	17366.4	99270.88	81904.48

2.2 Clustering with Fuzzy C-Means

Clustering is a technique for grouping data based on similarities between data in a dataset; this process is often called unsupervised learning. Some data belonging to the same cluster mean that the data have a higher level of similarity in that cluster compared to other clusters [7]. There are two clustering methods, namely hard clustering and fuzzy clustering. For hard clustering, each data point is only part of a cluster with a full membership value [24], whereas for fuzzy clustering, each data point allows being a member of more than one cluster with a membership value between 0 and 1 [25]. Therefore, the difference is the size of the membership value in a particular cluster. In 1981, Jim Bezdek proposed the FCM algorithm, one of the fuzzy clustering methods. The functioning of FCM is based on the number of clusters determined at the beginning by calculating the distance between each data point and each cluster centre [26]. First, there is an initialisation process for the membership of all data in each cluster. Then, the membership value of the data in each cluster is evaluated by repeatedly calculating the distance to the centre of the cluster until it finds the correct location. Calculating the centroid in each cluster uses a Equation 1 while the concept of distance used in this study uses seven distance concepts, including the Euclidean, Manhattan, Minkowski, Chebyshev, squared Euclidean, Canberra, and the combination Minkowski and Chebyshev distance.
The calculation for the centroid of the k-cluster is presented below:

V_{kj}, where k = 1, 2, …., c; and j = 1, 2, …, n

$$V_{kj} = \frac{\sum_{i=1}^{n}((U_{ik})^{w} X_{ij})}{\sum_{i=1}^{n}(U_{ik})^{w}} \quad (1)$$

The Euclidean distance is the concept most frequently used in several classification algorithms, namely k-NN, and clustering algorithms, such as k-means and FCM. Euclides stated that the Pythagorean metric could know the shortest distance between two points after this, called the Euclidean distance. The Euclidean distance d_2(x, y) between the points x and y, where x, y Rn, obtained from the calculations using Equation 2 indicates the number of data dimensions.

$$d_2(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (2)$$

The Manhattan distance was introduced by Hermann Minkowski in the late 19th century; the way it worked was to absolutely sum the difference of several Cartesian coordinates by Equation 3.

$$\mathrm{d_1(X,Y)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} |\mathrm{X_i} - \mathrm{Y_i}| \quad (3)$$

Minkowski uses the exponent p in his formulation. Obtain the Minkowski distance from the generalisation of the Euclidean (p = 2) and Manhattan (p = 1) distances based on Equation 4. The requirement for the metric conditions is that as long as the p-value is equal to or greater than 1, no p < 1 is allowed.

$$\mathrm{d_p\,(X,Y)} = \sqrt[\mathrm{p}]{\sum_{\mathrm{i=1}}^{\mathrm{n}} |\mathrm{X_i} - \mathrm{Y_i}|^{\mathrm{p}}} \quad (4)$$

When the p-value is an infinite positive number, the Chebyshev distance is obtained according to Equation 5.

$$d_\infty(X,Y) = \ _{i=1}^{n}\max |\, X_i - Y_i \,| \quad (5)$$

Four-distance metrics (Euclidean, Manhattan, Minkowski, and Chebyshev) are the most well-known and general basis for research.

The distance metric used in this case is the squared Euclidean distance (d_{SD}), also known as the sum of squared differences. It is a fundamental metric in most minor squares problems and linear algebra based on Equation 6.

$$d_{SD}(X,Y) = \sum_{i=1}^{n}(X_i - Y_i)^2 \quad (6)$$

Lance et al. have introduced the Canberra Range, which is a weighted version of the Manhattan distance since the 1960s using Equation 7.

$$d_\infty(X,Y) = \sum_{i=1}^{n}\frac{|X_i - Y_i|}{|X_i| + |Y_{i|}} \quad (7)$$

Rodrigue has introduced a new distance concept named the Minkowski and Chebyshev combination distance shown in Equation 8 for its acquisition.

$$d_{(w_1,w_2,p)}(x,y) = w_{\ 1}\sqrt[p]{\textstyle\sum_{k=1}^{n}|X_k - Y_k|^p} + \quad w_2 max_{k=1}^{n}|X_k - Y_k| \quad (8)$$

Here X_k and Y_k are the values of n-dimensional x and y, respectively, when the value of W_1 is greater than W_2, as is the case with Minkowski. Conversely, if W_2 is greater than W_1, it is the same as in Chebyshev. In general, Table 3 explains how the FCM algorithm works. First, calculate the value of the objective function in iteration (P) using Equation 9, while calculate the change in the partition matrix using Equation 10.

$$\mathrm{P} = \sum_{\mathrm{i=1}}^{\mathrm{n}}\sum_{\mathrm{k=1}}^{\mathrm{c}}(\mathrm{U_{ik}})^{\mathrm{w}}\ \mathrm{d_{ik}(X_i, V_k)} \quad (9)$$

$$U_{ik} = \frac{\left[\sum_{j=1}^{n} d_{ij}(X_i, V_j)\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{n} d_{ik}(X_i, V_k)\right]^{\frac{-1}{w-1}}} \quad (10)$$

Where i = 1,2,…., n ; and k = 1,2,…, c
Note: d_ik (X_i,V_k) The following formula can use any of the list distance metrics above.

Table 3. Fuzzy C-Means Algorithm

```
Algorithm 1: Fuzzy C-Means Clustering
BEGIN
        INPUT X, c, w, MaxIter, Epsilon, P0, t
        X ← Dataset to be clustered
        c ← The number of clusters
        w ← Rank
        MaxIter ← Maximum iterations
        Epsilon ← The smallest expected error
        P0, ← Initial objective function
        t ← Initial iteration
        i ← t
        While ( t < MaxIter ) do
                IF t value is equal to i value THEN
                        Initialize partition matrix U(t) randomly
                ELSE
                        Update Partition matrix U(t) with DistanceMetric(t)
                Update centroid V(t) with U(t)
                Calculate objective Function (P) with DistanceMetric(t) and U(t)
                IF | P- P0 | < epsilon THEN
                        Break
                ELSE
                     THEN
                        P0 ← P
                        t ← t + 1
        Return U*FCM ← U(t) and V*FCM  ← V(t)
END
```

2.3 Clustering with Modified Fuzzy C-Means Algorithm

In this study, we propose modifying the algorithm for fuzzy C-means in the random partition matrix initialisation process so that the cluster results obtained are inconsistent and always change if done repeatedly. The details of the proposed algorithm modification to initialise the partition matrix are presented in Table 4.

Table 4. Algorithm for Initialising the Partition

```
Algorithm 2: for Initialize partition matrix U(0)
BEGIN
     DN ← Normalize the input data X using MinMaxScaler
     U ← [ ]
     FOR k = 0 to the number of rows in the input data X
       FOR i = 0 to the number of clusters
         IF the value of i is equal to 0
             a = DN [k][i]
             U [k][i] = a
         ELSE IF the value of i is equal to (the number of clusters -1) Deduction = 0
              FOR x = 0 to (the number of clusters -1)
                        Deduction = Deduction + U[k][x]
              U[k][i] = 1-Deduction
         ELSE IF the value of i is greater than or equal to the number of columns of input
         data X
              b = 1 - a
              IF the value of b is less than or equal to DN [k] [i -number of columns of
             input data X]
                c = b
                a = a + c
                U[k][i]= c
             ELSE
                c = DN[k][i - number of columns of input data X]
                a = a + c
                U [k][i]= c
          ELSE
              b = 1 - a
              IF the value of b is less than or equal to DN [k][i]
```

```
                    c = b
                    a = a + c
                    U [k] [i] = c
                ELSE
                    c = DN[k][i]
                    a = a + c
                U [k][i] = c
            END FOR
        END FOR
END
```

2.4 Cluster Performance Evaluation

To assess the validity of the optimal number of clusters and explain the data structure, two clustering algorithms were used to measure the degree of compactness within each cluster and the separation between clusters. This study compares the performance of two clustering algorithms, namely the fuzzy C-Means algorithm and the modified fuzzy C-Means algorithm using three measures called PEI, PCI, and SS.

1) Partition Entropy Index (PEI).
   PEI is a measure that provides information about the membership matrix without considering the data itself. The minimum value implies a good partition in the sense of a sharper partition. The PEI value is obtained using Equation 11.

$$\mathrm{PEI} = -(\frac{1}{\mathrm{N}}) \sum_{c=1}^{\mathrm{C}} \sum_{i=1}^{\mathrm{N}} \mu_{ci} \log(\mu_{ci}) \tag{11}$$

   Generally, the optimal cluster is obtained if the value obtained is close to small (close to 0).

- Partition Coefficient Index (PCI).
  PCI is an index that measures partition fuzziness without considering the data set itself. PCI is a heuristic measure because it does not have a relation to the properties of the data. Therefore, its maximum value implies good partitioning in the least fuzzy clustering sense. The PCI value is obtained on the basis of Equation 12.

$$\mathrm{PCI} = (\frac{1}{\mathrm{N}}) \sum_{c=1}^{\mathrm{C}} \sum_{i=1}^{\mathrm{N}} \mu_{ci}^{2} \tag{12}$$

  The optimal cluster is based on the PCI value if the value obtained is more excellent (closer to 1) with a range of deals from 0 to 1.

- Silhouette Score (SS). The SS method combines two methods: the cohesion method, which measures how close the relationships are between objects in a cluster, and the separation method, which measures how far a set is from other groups. For SS, it is an optimal cluster if the value obtained is more excellent (close to 1) for the range of deals from -1 to 1 obtained based on Equation 13.

$$SS = (b - a)/\max(a, b) \tag{13}$$

  Where, a= average intracluster distance, i.e. the average distance between each point within a cluster, and b= average intercluster distance, i.e. the average distance between all clusters.

3. Results and Discussion

This study carried out experiments by clustering DO and TDS data. The aim is to determine the carrying capacity of river pollution loads using a modified FCM clustering algorithm with several clusters, namely 2, 3, and 4 using six distances (Euclidean, Canberra, Chebyshev, Manhattan, Minkowski, and Squared). The result of the three parameter validation tests, namely PCI, PEI and SS, shows that 2 clusters have the most optimal clustering results. The detailed results are presented in Figure 1 and Figure 2. For example, the results of the DO data cluster using the Euclidean distance with 2 clusters have optimal performance with PCI, PEI, and SS values of 0.68, 0.71, and 0.58 respectively. Similar results are obtained with 2 clusters using the Canberra, Chebyshev, Manhattan, Minkowski and squared Euclidean distance.

The following experiment was to determine the effect on clustering results using several different distances. The clustering algorithm that used both FCM and FCM modifications had the same clustering results when measured using all three parameters, namely PEI, PCI, and SS. For example, the clustering results using Minkowski distance (p=0.5) and Chebyshev distance (W1=1 and W2=2), resulted the FCM algorithm with PEI, PCI, and SS performance of 0.817, 0.618, and 0.513 respectively. However, using the modified FCM algorithm resulted PEI, PCI, and SS of 0.816, 0.618 and 0.513 respectively; detailed data are reported in Table 5. The FCM algorithm produced the best clustering results,

and the modified FCM using squared Euclidean distance produced PE, PCI and SS of 0.287, 0.889, and 0.681 respectively.

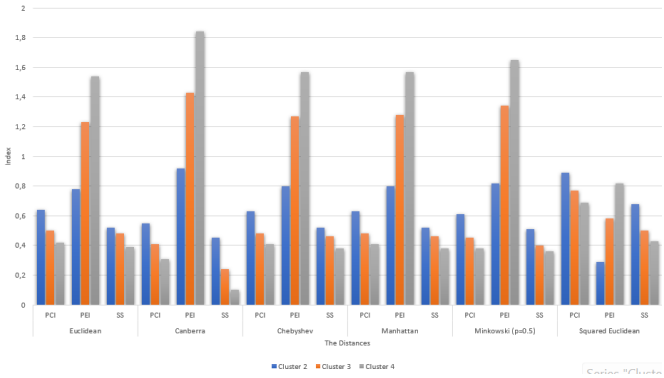

Figure 1. The results of DO Data Clustering using Six Distances

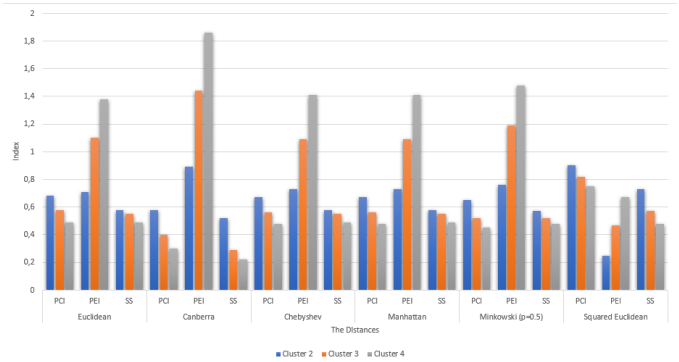

Figure 2. The Results of TDS Data Clustering using Six Distances

Table 5. Clustering Results using the FCM and Modified FCM Algorithm

Distance metrics	FCM			Modified FCM		
	PEI	PCI	SS	PEI	PCI	SS
Minkowski (p = 0.5) & Chebyshev (W1 = 1, W2 = 2)	0.817	0.618	0.513	0.816	0.618	0.513
Minkowski (p = 0.75) & Chebyshev (W1 = 1, W2 = 2)	0.804	0.625	0.519	0.806	0.624	0.519
Minkowski (p = 1) & Chebyshev (W1 = 1, W2 = 2)	0.798	0.629	0.519	0.798	0.629	0.519
Minkowski (p = 2) & Chebyshev (W1 = 2, W2 = 1)	0.787	0.635	0.519	0.787	0.635	0.519
Minkowski (p = 3) & Chebyshev (W1 = 2, W2 = 1)	0.786	0.635	0.519	0.786	0.635	0.519
Minkowski (p = 4) & Chebyshev (W1 = 2, W2 = 1)	0.788	0.635	0.519	0.788	0.635	0.519
Euclidean	0.783	0.637	0.519	0.783	0.637	0.519
Canberra	0.923	0.551	0.450	0.923	0.551	0.450
Chebyshev	0.798	0.629	0.519	0.798	0.629	0.519
Manhattan	0.798	0.628	0.519	0.798	0.628	0.519
Minkowski (p = 0.5)	0.822	0.614	0.513	0.822	0.614	0.513
Minkowski (p = 0.75)	0.808	0.623	0.519	0.808	0.623	0.519
Minkowski (p = 3)	0.782	0.638	0.519	0.782	0.638	0.519
Minkowski (p = 4)	0.783	0.637	0.519	0.783	0.637	0.519
Squared Euclidean	**0.287**	**0.889**	**0.681**	**0.287**	**0.889**	**0.681**

Initial random matrix initialisation, as in the FCM algorithm or not done randomly, as the algorithm proposed in this study, which is called the modified FCM algorithm, does not affect the performance of the clustering results. However, several experiments show that the initial partition matrix initialisation affects the number of iterations. Using the non-random initial partition matrix initialisation (modified FCM algorithm) results in fewer iteration stops than random matrix initialisation (FCM). In addition, using random initialisation causes each class's cluster centre to change. For example, using the Euclidean distance to cluster DO data with the FCM algorithm and the modified FCM algorithm produces different iteration stops. Based on three tests with random initial partition matrix initialisation, the iteration

stops varied, namely 31, 26, and 25, while using the modified FCM algorithm produced consistent and more minor iteration stops, namely 19. More details are shown in Table 6.

Table 6. Stop Iterating Clustering Results using the Fuzzy C-Means Algorithm and Modified Fuzzy C-Means

Distance metrics	Stop Iteration – DO				Stop Iteration – TDS			
	Random			Proposed	Random			Proposed
	1	2	3		1	2	3	
Euclidean	31	26	25	19	31	28	30	20
Canberra	38	44	37	28	33	27	27	21
Chebyshev	24	26	27	20	28	25	27	22
Manhattan	25	24	28	18	25	24	27	26
Minkowski (p = 0.5)	29	27	34	21	25	21	22	20
Minkowski (p = 0.75)	33	27	27	21	33	27	26	18
Minkowski (p = 3)	24	23	24	19	28	26	26	22
Minkowski (p = 4)	26	26	30	19	29	31	28	20
Squared Euclidean	39	39	39	31	31	32	32	24
Minkowski (p=0,5) & Chebyshev (w1=1,w2=1)	23	24	34	19	24	20	22	19
Minkowski (p=0,75) & Chebyshev (w1=1,w2=1)	19	21	23	21	30	28	27	20
Minkowski (p=1) & Chebyshev (w1=1,w2=1)	28	37	29	19	25	23	23	22
Minkowski (p=2) & Chebyshev (w1=1,w2=1)	21	24	29	21	21	29	23	22
Minkowski (p=3) & Chebyshev (w1=1,w2=1)	27	22	24	23	31	23	25	22
Minkowski (p=4) & Chebyshev (w1=1,w2=1)	27	23	23	21	24	33	25	24

The purpose of proposing modifications to the clustering algorithm focusses more on solving two problems that are often found in clustering algorithms, namely: random dataset initialisation and inconsistent number of iterations. Some previous research proposals include: proposed modifications to the k-means and k-mode algorithms, which proved to be able to overcome the randomness in the initialisation of the dataset matrix. The measurement results using the Davies-Bouldin index (DBI) and Silhouette Index (SI) performance of the k-mode modification algorithm are better than the modification algorithm on k-means. In this study only one distance concept, namely euclidean distance [31]. While optimisation of the fuzzy c-means algorithm using the concept of distance combination between Minkowski and Chebyshev distance and principal component analysis. The PCA technique is used for preprocessing by reducing data before calculating the distance from each data set to the cluster centre. Research has shown better results when compared to other distance concepts [23].

4. Conclusion

In this paper, we propose an algorithm to initialise the partition matrix for the FCM clustering algorithm called the modified FCM algorithm. The modified FCM algorithm with several distances has been used in the DO and DTS data clustering process to determine the carrying capacity of river pollution loads with varied number of clusters, namely 2, 3, and 4. The validation results of the PCI, PEI, and SS values show the number of clusters using various distances, i.e. 2 clusters have the best results. The performance of the two algorithms obtained the same validation results for all distances, but the stop iterations using FCM algorithm change every trial and tend to have more iteration stops. In contrast, the modified FCM algorithm has fewer iteration stops. The clustering process can be carried out for further research using the proposed algorithm/modified FCM for the other water quality data.

Acknowledgements

The authors gratefully acknowledge the sponsorship of the institute for research and community service UIN Sunan Kalijaga Yogyakarta based on SK No: 2096.16/Un.02/L3/TL/06/2022.

References

[1] Pusat Pengolahan Data Kementerian Pekerjaan Umum Republik Indonesia, *Buku Informasi Statistik Pekerjaan Umum*. 2012.
[2] Menteri LIngkungan Hidup dan Kehutanan Republik Indonesia, “Peraturan Menteri Lingkungan Hidup dan Kehutanan Republik Indonesia Nomor 10 Tahun 2022.”
[3] T. H. M. van Emmerik, S. Kirschke, L. J. Schreyers, S. Nath, C. Schmidt, and K. Wendt-Potthoff, “Estimating plastic pollution in rivers through harmonized monitoring strategies,” *Mar Pollut Bull*, vol. 196, Nov. 2023. https://doi.org/10.1016/j.marpolbul.2023.115503
[4] T. Garg, S. E. Hamilton, J. P. Hochard, E. P. Kresch, and J. Talbot, “(Not so) gently down the stream: River pollution and health in Indonesia,” *J Environ Econ Manage*, vol. 92, pp. 35–53, Nov. 2018. https://doi.org/10.1016/j.jeem.2018.08.011
[5] Z. Feng, R. Zhang, X. Liu, Q. Peng, and L. Wang, “Agricultural nonpoint source pollutant loads into water bodies in a typical basin in the middle reach of the Yangtze River,” *Ecotoxicol Environ Saf*, vol. 268, Dec. 2023. https://doi.org/10.1016/j.ecoenv.2023.115728
[6] C. Team *et al.*, *Fresh Water for the future*, June. United Nations Environment Programme, 2012.
[7] A. Development Bank, “ADB Annual Report 2016,” 2016.

[8] D. Sutjiningsih, "Water Quality Index for Determining the Development Threshold of Urbanized Catchment Area in Indonesia," *International Journal of Technology*, vol. 8, no. 1, p. 143, 2017. https://doi.org/10.14716/ijtech.v8i1.3971
[9] Y. Li *et al.*, "Study on total phosphorus pollution load estimation and prevention and control countermeasures in Dongting Lake," *Energy Reports*, vol. 9, pp. 294–305, Aug. 2023. https://doi.org/10.1016/j.egyr.2023.04.272
[10] K. K. Verma, B. M. Singh, and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system," *International Journal of Information Technology (Singapore)*, vol. 14, no. 1, pp. 397–410, Feb. 2022. https://doi.org/10.1007/s41870-019-00364-0
[11] Á. López-Oriona, J. A. Vilar, and P. D'Urso, "Hard and soft clustering of categorical time series based on two novel distances with an application to biological sequences," *Inf Sci (N Y)*, vol. 624, pp. 467–492, May 2023. https://doi.org/10.1016/j.ins.2022.12.065
[12] Á. López-Oriona, P. D'Urso, J. A. Vilar, and B. Lafuente-Rego, "Quantile-based fuzzy C-means clustering of multivariate time series: Robust techniques," *International Journal of Approximate Reasoning*, vol. 150, pp. 55–82, Nov. 2022. https://doi.org/10.1016/j.ijar.2022.07.010
[13] L. Zhu, "Selection of Multi-Level Deep Features via Spearman Rank Correlation for Synthetic Aperture Radar Target Recognition Using Decision Fusion," *IEEE Access*, vol. 8, 2020. https://doi.org/10.1109/ACCESS.2020.3010969
[14] S. Subudhi and S. Panigrahi, "Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 5, 2020. https://doi.org/10.1016/j.jksuci.2017.09.010
[15] N. Jafarzade *et al.*, "Viability of two adaptive fuzzy systems based on fuzzy c means and subtractive clustering methods for modeling Cadmium in groundwater resources," *Heliyon*, vol. 9, no. 8, Aug. 2023. https://doi.org/10.1016/j.heliyon.2023.e18415
[16] A. Gupta, S. Datta, and S. Das, "Fuzzy Clustering to Identify Clusters at Different Levels of Fuzziness: An Evolutionary Multiobjective Optimization Approach," *IEEE Trans Cybern*, vol. 51, no. 5, pp. 2601–2611, May 2021. https://doi.org/10.1109/TCYB.2019.2907002
[17] A. S. Shirkhorshidi, T. Y. Wah, S. M. R. Shirkhorshidi, and S. Aghabozorgi, "Evolving Fuzzy Clustering Approach: An Epoch Clustering That Enables Heuristic Postpruning," *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 3, pp. 560–568, Mar. 2021. https://doi.org/10.1109/TFUZZ.2019.2956900
[18] J. E. Nalavade and T. Senthil Murugan, "HRNeuro-fuzzy: Adapting neuro-fuzzy classifier for recurring concept drift of evolving data streams using rough set theory and holoentropy," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 498–509, Oct. 2018. https://doi.org/10.1016/j.jksuci.2016.11.005
[19] W. Yiping *et al.*, "An improved multi-view collaborative fuzzy C-means clustering algorithm and its application in overseas oil and gas exploration," *J Pet Sci Eng*, vol. 197, Feb. 2021. https://doi.org/10.1016/j.petrol.2020.108093
[20] G. Wang, S. Guo, L. Han, Z. Zhao, and X. Song, "COVID-19 ground-glass opacity segmentation based on fuzzy c-means clustering and improved random walk algorithm," *Biomed Signal Process Control*, vol. 79, Jan. 2023. https://doi.org/10.1016/j.bspc.2022.104159
[21] Y. Huang, D. Chen, W. Zhao, and Y. Lv, "Fuzzy C-Means Clustering Based Deep Patch Learning With Improved Interpretability for Classification Problems," *IEEE Access*, vol. 10, pp. 49873–49891, 2022. https://doi.org/10.1109/ACCESS.2022.3171109
[22] H. Murfi, N. Rosaline, and N. Hariadi, "Deep autoencoder-based fuzzy c-means for topic detection," *Array*, vol. 13, p. 100124, Mar. 2022. https://doi.org/10.1016/j.array.2021.100124
[23] S. Surono and R. D. A. Putri, "Optimization of Fuzzy C-Means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis," *International Journal of Fuzzy Systems*, vol. 23, no. 1, pp. 139–144, Feb. 2021. https://doi.org/10.1007/s40815-020-00997-5
[24] É. O. Rodrigues, "Combining Minkowski and Chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier," 2018. https://doi.org/10.1016/j.patrec.2018.03.021
[25] M. S. H. Ardani *et al.*, "A new approach to signal filtering method using K-means clustering and distance-based Kalman filtering," *Sens Biosensing Res*, vol. 38, Dec. 2022. https://doi.org/10.1016/j.sbsr.2022.100539
[26] Á. López-Oriona, J. A. Vilar, and P. D'Urso, "Quantile-based fuzzy clustering of multivariate time series in the frequency domain," *Fuzzy Sets Syst*, vol. 443, pp. 115–154, Aug. 2022. https://doi.org/10.1016/j.fss.2022.02.015
[27] F. Farid and D. Rosadi, "Portfolio optimization based on self-organizing maps clustering and genetics algorithm," *International Journal of Advances in Intelligent Informatics*, vol. 8, no. 1, pp. 33–44, Mar. 2022. https://doi.org/10.26555/ijain.v8i1.587
[28] A. E. Haryati, S. Surono, and S. Suparman, "Implementation of Minkowski-Chebyshev Distance in Fuzzy Subtractive Clustering," *EKSAKTA: Journal of Sciences and Data Analysis*, pp. 82–87, Jun. 2021. https://doi.org/10.20885/eksakta.vol2.iss2.art1
[29] J. Li, J. Shao, W. Wang, and W. Xie, "An evolutional deep learning method based on multi-feature fusion for fault diagnosis in sucker rod pumping system," *Alexandria Engineering Journal*, vol. 66, pp. 343–355, Mar. 2023. https://doi.org/10.1016/j.aej.2022.11.028
[30] N. Gueorguieva, I. Valova, and G. Georgiev, "M&MFCM: Fuzzy C-means Clustering with Mahalanobis and Minkowski Distance Metrics," in *Procedia Computer Science*, Elsevier B.V., 2017, pp. 224–233. https://doi.org/10.1016/j.procs.2017.09.064
[31] E. Setyaningsih, N. Hidayat, U. Lestari, and A. Septiarini, "Modification OF K-Means and K-Mode Algorithms to Enhance the Performance of Clustering Student Learning Styles in the Learning Management System," *ICIC Express Letters*, vol. 17, no. 1, pp. 49–59, Jan. 2023. https://doi.org/10.24507/icicel.17.01.49