Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2021.26696

213 (2021) 84–90
February

Water quality assessment based on interval-valued data cluster analysis

Siqing Shan[a], Yuebin Bai[b,*], Xiaojing Wang[b]

[a]*School of Economic and Management, Beihang University, Beijing 100191, China, email: shansiqing@buaa.edu.cn (S. Shan)*
[b]*School of Computer Science and Engineering, Beihang University, Beijing 100191, China, Tel. +13810972580; emails: byb@buaa.edu.cn (Y. Bai), 317745556@qq.com (X. Wang)*

Received 17 January 2020; Accepted 24 September 2020

ABSTRACT

The interval data cluster analysis method was adopted to evaluate the water quality of the Huaihe River. This method could reduce the dimension of water quality data. When the dimension reduced, the calculation could be easier. With this method, there is no loss of information. Twenty-six sites in the river were selected to take samples. At each sampling site, four indicators were recorded each week in 2012. First, the original 1,326 × 4 matrix was converted into a 26 × 4 matrix. The traditional number elements in matrix were replaced by interval-valued data. Then, the interval data were standardized. For the standardized data, Euclidean–Hausdorff distance was used for hierarchical cluster analysis. To determine the needed clusters, the paper employed corrected rand index (CRI). According to the value of CRI, 26 sites were divided into six clusters. Samples in different clusters had different interval radius and pollution levels. Samples in cluster 1 are mildly polluted and the fluctuation of indicators' concentrations is not as violent as those in other clusters. Though the samples in cluster 2 and 4 rank in all the middle levels in terms of pollution, pollution in cluster 4 is relatively stable. Cluster 5 and 6 have higher concentrations of DO, COD_{Mn}, and NH_3–N, due to insufficient DO. The interval data cluster analysis method based on Euclidean–Hausdorff distance classifies the sample sites into multiple clusters without loss of information.

Keywords: Interval-valued data; Water quality assessment; Cluster analysis; Euclidean-Hausdorff distance; Corrected rand index

1. Introduction

In recent years, with the deterioration of water quality, water pollution has become serious and water quality assessment has been a heated topic. Many methods have been developed to evaluate water pollution. Entropy method is used to calculate the weight during the evaluation process [1]. Multivariate statistics analysis is very useful for water quality assessment [2–7]. Fuzzy set was first proposed by Zadeh [8]. It has been applied in many fields. This method is used to assess water quality combined with multivariate statistics [9,10]. Neural network is also a very popular method in water quality assessment. BP neural network is applied to water quality assessment for Miyun Reservoir recharged with reclaimed water [11]. He et al. [12] proposed an Radial Basis Function neural network based on optimized parameters of genetic algorithm in terms of water quality evaluation.

The diversification of data information promotes the demand for efficient data analysis methods. When the data structure is too complicated, the traditional data analysis technology has many limitations. The sample sizes and variable dimensions are always large, so it takes too much time to calculate the cost. This also makes it difficult to maintain the internal relationship of data attributes and cannot obtain the implied knowledge in the data [13]. At the same

* Corresponding author.

time, the amount of water quality data is large, and the above methods may lose some information when evaluating the degree of pollution. Traditional methods should be optimized for evaluating these complex water quality data. Interval-valued data analysis method is an effective way for processing high-dimension data. Interval analysis, also known as interval mathematics, is a branch of mathematics in which interval variables replace point variables. It was originally developed from the error theory of computational mathematics.

It is a symbolic data analysis (SDA) method. Symbolic data analysis is defined as a theory to extract systematic knowledge from massive data and researches on how to explore the system theory and method from massive data [14]. SDA can reduce the computational complexity and use the data packaging method to maintain the characteristics of the sample. After packing, the samples are called symbolic objects and the numbers become symbolic data. Symbolic data can be both quantitative or qualitative. Interval-valued data is a common type of symbolic data [15]. For example, the concentration of BOD in a river may vary in the interval [3.6, 10.8] (mg/L). For interval-valued water quality data, cluster analysis (CA) could be used to classify similar sample sites or indicators. Cluster analysis mainly contains hierarchical and partitional clustering [16,17]. It is a data analysis field still under exploration. CA aims to organize a set of items into several clusters so that items within a given cluster have a high degree of similarity, while items belonging to different clusters are greatly different. There are also many studies using multivariate analysis methods to cope with water quality data [18–22].

When point data are converted into symbolic data, the traditional cluster analysis methods become invalid. Therefore, it is necessary to improve the cluster analysis methods so that these methods could deal with symbolic data [23–25]. In recent years, many symbolic data cluster analysis methods have been put forward. Transformation algorithm could be used for clustering of distributed symbolic variables [26]. Similarity measures about data decomposition rules and cluster methods were proposed for multi-valued constraint symbolic data [27]. Due to the important application of interval value data, scholars have conducted a lot of researched on cluster analysis on interval-valued data cluster analysis. For the objects described by interval data, a partition clustering method based on dynamic clustering and *L*2 distance (Euclidean distance) is introduced [28]. A fuzzy clustering algorithm based on Mahalanobis distance was proposed for partitioning symbolic interval data [29]. A partitioning fuzzy *K*-means clustering model for interval-valued data was presented with suitable adaptive quadratic distance. In addition, additional interpretation tools, which were suitable for these fuzzy clustering models were put forward for individual fuzzy clusters of interval-valued data, [30]. In order to compare symbolic interval data, researchers also introduced dynamic cluster methods for partitioning symbolic interval data based on city-block distance and Hausdorff distance [31]. Wasserstein-based distance generalized a wide set of distances for interval data based on different approaches or in different contexts of analysis when used to clustering techniques for interval-valued data [32]. Fuzzy clustering method and kernel function were integrated into a clustering algorithm for interval numbers so as to handle the problem in the existing similar clustering algorithms for multi-pattern prototypes and asymmetric data structure [33]. Aiming at the problem of multi-attribute clustering analysis with uncertain interval numbers, a clustering algorithm with numerical data as the clustering center was proposed [34]. Li et al. [35] defined the general distributed interval data and gave the hierarchical cluster method based on Euclidean–Hausdorff distance.

Until now, a few researchers have been interested in multivariate statistics of interval water quality data. This method can deal with high dimensional data. Because water quality data always contain many sites and indicators for a long period, this method may have a positive effect on water quality assessment [36].

Huaihe River is one of three major rivers in China with a large population and rapid economic development. The Huaihe River, featured by surface water, is located in the south part of China. The source of the Huaihe River comes from the north of Tongbai Mountain in Henan Province, China. The basin contains five provinces, Hubei, Henan, Anhui, Shandong, and Jiangsu Provinces. The total population is 165 million and the total area is 270 thousand square meters in the basin. This river provides water for such a large area, so its water quality should be given much more attention. The data given in the paper are obtained from the Ministry of Ecology and Environment of the People's Republic of China.

The structure of this paper is as follows. The second part introduces the interval-valued data cluster analysis methods and research objects. The third part is about the use of these methods on water quality evaluation in Huaihe River, China.

2. Methods and theories

2.1. Interval-valued data

Interval-valued data always means that the values of samples are not determined by numerical values. Instead, these values may contain a particular range of data on the set of real numbers. Interval-valued data could be expressed as $x=\left\{t \mid \underline{x} \le t \le \overline{x}, \underline{x}, \overline{x} \in \mathbb{R}, \overline{x} \le \overline{x}\right\}$. Where $\underline{x}$ and $\overline{x}$ are the lower and upper bounds of the interval data respectively. When $\underline{x}=\overline{x}$, the interval data become numeric ones. In addition, interval-valued data could also be expressed in the other two forms. The first one is the ordered array that contains both lower bound and upper bound, $x=[\underline{x}, \overline{x}]$. The second form is an array containing the center and radius of the interval-valued data, $x = (x^c, x^r)$. $x^c=\frac{1}{2}\left(\underline{x}+\overline{x}\right)$ is the center of the interval and $x^r=\frac{1}{2}\left(\overline{x}-\underline{x}\right)$ means the radius.

For n-dimensional vector $X = (x_1,x_2,\ldots,x_n)'$, if all elements of the vector are interval-valued data, that is $x_i=\left[\underline{x}_i \le \overline{x}_i\right]\left(1 \le i \le n\right)$, then X is an n-dimensional interval-valued vector. For $n \times P$-dimensional matrix $X_{n\times p}=(x_{ij})_{n\times p}$, if all elements are interval-valued data, $x_{ij}=\left[\underline{x}_{ij}, \overline{x}_{ij}\right]\left(1 \le i \le n, 1 \le j \le p\right)$, then $X_{n\times p}$ is expressed as an $n \times p$-dimensional interval-valued data matrix. The multivariate statistical analysis for water quality data is an interval-valued data table containing samples and variables. This table is an $n \times p$-dimensional interval-valued matrix.

$$X_{n\times p}=\begin{pmatrix}\left[\underline{x}_{11},\bar{x}_{11}\right] & \left[\underline{x}_{12},\bar{x}_{12}\right] & \cdots & \left[\underline{x}_{1p},\bar{x}_{1p}\right]\\ \left[\underline{x}_{21},\bar{x}_{21}\right] & \left[\underline{x}_{22},\bar{x}_{22}\right] & \cdots & \left[\underline{x}_{2p},\bar{x}_{2p}\right]\\ \vdots & & & \\ \left[\underline{x}_{n1},\bar{x}_{n1}\right] & \left[\underline{x}_{n2},\bar{x}_{n2}\right] & \cdots & \left[\underline{x}_{np},\bar{x}_{np}\right]\end{pmatrix}=\begin{pmatrix}e_1'\\ e_2'\\ \vdots\\ e_n'\end{pmatrix}=\left(X_1\ X_2\cdots X_p\right) \tag{1}$$

In each line in the matrix, e_i ($1 \le i \le n$), refers to an interval sample and each column, X_j ($1 \le j \le p$), is an interval variable. Interval sample e_i is expressed by p interval variables and interval variable X_j contains n independent observations. Each observation is an interval data.

2.2. Hierarchical clustering analysis

Hierarchical cluster analysis is a cluster analysis method that attempts to establish clustering levels. There are two types of strategies to conduct CA, agglomerative and divisive. The agglomerative strategy was adopted. In this method, each observation starts in its own cluster, and when moving up the hierarchy, the cluster pairs merge into one cluster pair. The two clusters with the smallest distance are merged into one cluster at a time. When all samples are merged into the same cluster, the calculation will end. When performing hierarchical cluster analysis for interval-valued data, traditional method should be improved. The procedures for interval-valued data CA are as follows.

First, the data should be standardized. For a specific interval variable X_j, the standardization of its sample observation value $X_{kj} = [a_{kj}, b_{kj}](k = 1,\ldots,n)$ follows that of traditional point data. Set U_{kj} as the standardized variable of X_{kj}, then:

$$U_{kj}=\frac{X_{kj}-\bar{X}_j}{S_j}=\left[\frac{a_{kj}-\bar{X}_j}{S_j},\frac{b_{kj}-\bar{X}_j}{S_j}\right]=\left[c_{kj},d_{kj}\right] \tag{2}$$

where $\bar{X}_j$ is the mean of interval variable X_j, S_j is the standard deviation of interval variable X_j, the calculation of $\bar{X}_j$ and S_j are as follows:

$$\bar{X}_j=\int_{-\infty}^{+\infty}xf_{x_j}(x)dx=\frac{1}{n}\sum\nolimits_{k=1}^{n}\mu_{kj} \tag{3}$$

$$S_j^2=\frac{1}{n}\sum_{k=1}^{n}\left(\sigma_{kj}^2+\left(\bar{X}_j-\mu_{kj}\right)^2\right) \tag{4}$$

where μ_{kj} is the mean of general distribution interval variable X_{kj}. As it is difficult to obtain μ_{kj}, it always uses the sample mean $\bar{X}_{kj}$ as the estimator of μ_{kj}. σ_{kj}^2 is the variance of general distribution interval variable X_{kj}. The sample variance S_{kj}^2 is always used as estimator for σ_{kj}^2.

Secondly, it is to calculate the distances between every two samples and obtain the distance matrix $D^{(0)}$. Hausdorff distance is used and the Hausdorff distance of two interval numbers is calculated as: $H(A,B)=\left|c(A)-c(B)\right|+\left|r(A)-r(B)\right|$, $c(X)=\frac{m+n}{2}$, and $r(X)=\frac{n-m}{2}$ are the center and radius of interval number $X = [m,n]$, respectively.

For interval variables, let $X = (x_1,x_2,\ldots,x_p)^T = ([a_1,b_1], [a_2,b_2],\ldots,[a_p,b_p])^T$ and $Y = (y_1,y_2,\ldots,y_p)^T = ([c_1,d_1],[c_2,d_2],\ldots, [c_p,d_p])^T$.

where X and Y are both interval vectors that contain p variables. The Euclidean–Hausdorff distance in n-dimensional real space is extended into two interval variables:

$$d_H(X,Y)=\sqrt{\sum_{i=1}^{p}\left(\left|c(x_i)-c(y_i)\right|+\left|r(x_i)-r(y_i)\right|\right)^2} \tag{5}$$

The distance between two compact sets in $\Re^p$ space is always expressed as Euclidean–Hausdorff distance. When two interval vectors are merged as one cluster, the new cluster, also an interval vector, is still a compact set, so the distance between two clusters can be expressed by Euclidean–Hausdorff distance.

Thirdly, set s as the number of iterations, k as the number of clusters, $s = 1$ and $k = n$. Each sample is a cluster, the i-th cluster is $G_i\{x_i\}(i = 1,\ldots,n)$. The distances between two clusters refer to the distances between two samples.

Fourthly, according to the distance matrix $D^{(s)}$, two clusters with the minimum distance are merged. For two interval vectors clusters X and Y, the new merged cluster Z can also be expressed by interval vector:

$$\begin{aligned}Z&=(z_1,z_2,\cdots,z_p)^T\\ &=([\min(a_1,c_1),\max(b_1,d_1)],[\min(a_2,c_2),\max(b_2,d_2)],\\ &\quad\cdots,[\min(a_p,c_p),\max(b_p,d_p)])^T\end{aligned} \tag{6}$$

Let $k = n–s$, if $k > 1$, move to step five, otherwise, to step six.

Fifthly, continuing the next calculation. Let $s = s + 1$ and calculate the distance between new cluster and the others. According to the distance matrix $D^{(s)}$, turn to step four.

Sixthly, $k = 1$ and all the clusters are merged as one, the calculation ends.

2.3. Corrected rand index

In order to determine the number of needed clusters, the corrected rand index (CRI) was studied. Given an n object set $S = \{O_1,\ldots,O_n\}$, suppose $U = \{u_1,\ldots,u_R\}$, and $V = \{v_1,\ldots,v_C\}$ represent two different partitions of S, that is, the entries in U and V are subsets of S; $\bigcup_{i=1}^{R}u_i=S=\bigcup_{j=1}^{C}u_j$; $u_i\bigcap u_{i'}=v_j\bigcap v_{j'}$ for $1 \le i \ne i' \le R$ and $1 \le j \ne j' \le C$. Take n_{ij} to denote the number of objects that are common to classes u_i and v_j, the information that overlaps between the two partitions U and V can be expressed in the form of contingency tables (using standard "dot" notation for row and column sums) with n_i and n_j referring respectively to the number of objects in classes u_i (row i) and v_j (column j), as shown in Table 1. A binomial coefficient $\binom{m}{2}$ is defined as 0 when $m = 0$ or 1. The CRI can be expressed:

Table 1
Notation for comparing two partitions

	Partition V					
Partition U	Class	v_1	v_2	⋯	v_C	Sums
	u_1	n_{11}	n_{12}	⋯	n_{1C}	n_1
	u_2	n_{21}	n_{22}		n_{2C}	n_2
	⋮	⋮	⋮		⋮	⋮
	u_R	n_{R1}	n_{R2}	⋯	n_{RC}	n_R
	Sums	n_1	n_2	⋯	n_C	$n = n$

$$\frac{\sum_{i,j}\binom{n_{ij}}{2}-\sum_{i}\binom{n_i}{2}\sum_{j}\binom{n_j}{2}\Big/\binom{n_i}{2}}{\frac{1}{2}\left[\sum_{i}\binom{n_i}{2}+\sum_{j}\binom{n_j}{2}\right]-\sum_{i}\binom{n_i}{2}\sum_{j}\binom{n_j}{2}\Big/\binom{n_i}{2}} \tag{7}$$

where n_{ij} refers to the same samples in cluster u_i and cluster v_j, n_i means the number of samples in cluster u_i and n_j is the number of samples in clusters v_j.

3. Results and discussions

According to the method described above, researchers analyzed based on the interval-valued data cluster method for Huaihe River. There were 26 monitoring points. The monitoring points were chosen from the source of the river to the estuary. The data were obtained weekly in 2012. As there was no record in the 16th week, there were data for 51 weeks in total. Four indicators were recorded, including pH, NH_3–N, DO, and COD_{Mn}. DO means a profitable index. The higher the concentration of DO is, the better the water quality will be. NH_3–N and COD_{Mn} are cost indicators. The average value of NH_3–N and COD_{Mn} is negatively correlated with that of DO. The lower the concentration is, the better the water quality will be. pH is an indicator which is used for the acidity and alkaline.

Due to the large data matrix, it is difficult to solve this problem using traditional multivariate statistical methods, so dimensional reduction is required. The interval-valued data technology actually was adopted to simplify the data. Finally, the 1,326 × 4 water quality data matrix were changed into a 26 × 4 matrix. Each element of the simplified matrix contained 51 data and they were within a specific interval. Table 2 shows the interval-valued data of the sampling sites. The 26 × 4 water quality data matrix was studied. This matrix contains all the information reflecting the pollution degree of the river.

The results of CA for these samples have been shown in Fig. 1. From Fig. 1, we can find that the samples in the same cluster have no rules.

The 26 monitoring sites could be classified into several clusters according to the distance. For example, at the distance of $0.56 < D < 0.75$, there are five clusters and when the distance is $0.75 < D < 1.02$, there are four clusters. For different distances, the numbers of clusters are different.

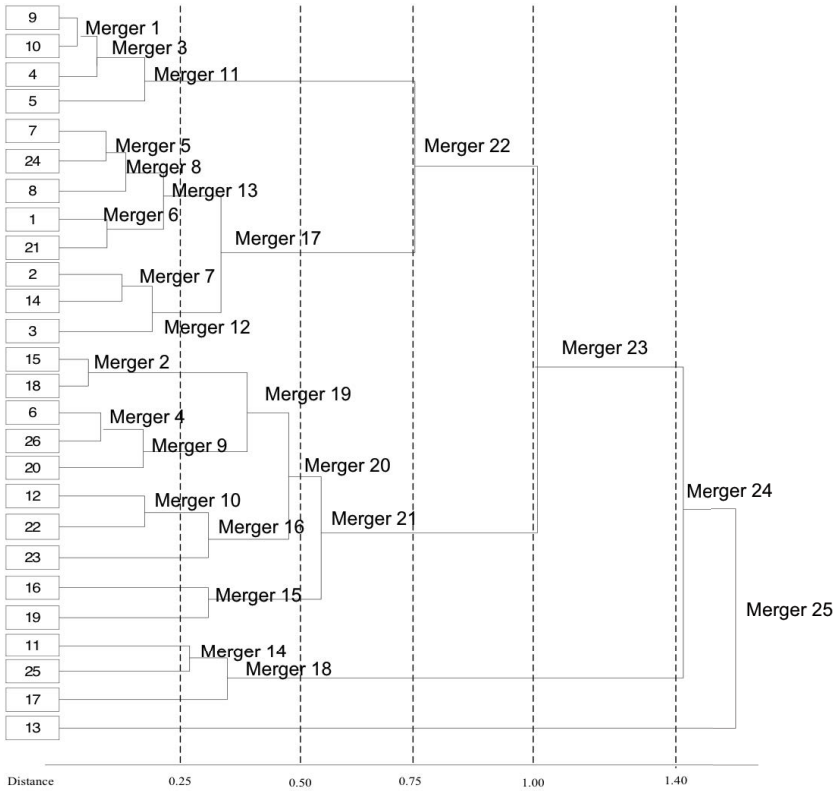

Fig. 1. Distance between the two smallest clusters for each calculation.

Therefore, researchers calculate the CRI from three clusters to seven clusters. The results are shown in Table 3. As the CRI between 5 and 6 clusters is bigger than others and CRI between 6 and 7 clusters is bigger than that between 4 and 5 clusters, it is proper to classify the samples into six clusters. The six clusters are shown below. Cluster 1 has eight samples, from site 1, 2, 3, 7, 8, 14, 21, and 24. Cluster 2 has four samples from site 4, 5, 9, and 10. Cluster 3 only contains that from site 13. Cluster 4 has three samples from site 11, 17, and 25. Cluster 5 contains eight samples, each from site 6, 12, 15, 18, 20, 22, 23, and 26. Site 16 and 19 are in cluster 6.

Table 4 shows the mean of interval-valued data for each cluster and Table 5 shows the radius of each cluster for the four indicators. These tables reflect the range of variation for six clusters. The indicator pH ranges around 8, so the water is alkalinity. For cluster 1, the concentration of DO, COD_{Mn}, and NH_3–N are not very dramatic. The samples in this cluster are less violent than in other clusters. The concentrations of COD_{Mn} and NH_3–N are also a little lower than those in other clusters. The concentration of DO is no different from other star clusters. This illustrates that samples in this cluster are less polluted and the water quality does not fluctuate a lot. For cluster 2 and 4, the pollution level of all indicators ranks in the middle level. Additionally, comparing with cluster 2, 4 is relatively stable. However, the concentration of DO is slightly higher than that in the cluster 4, while concentrations of COD_{Mn} and NH_3–N are lower than in cluster 2. Compared with cluster 4, the pollution in cluster 2 is proven to be lighter. In cluster 3, the concentration of DO is high and the concentrations of COD_{Mn} and NH_3–N are relatively low, so samples in this cluster incline to recovery in a much stronger manner. For cluster 5 and 6, the concentration of DO is also high, but the concentrations of COD_{Mn} and NH_3–N are much higher than those in other clusters. For these two clusters, the supply of DO cannot satisfy the demand, so the pollution of these samples is more serious. The last three indicators in cluster 6 fluctuate violently and no cluster has the maximum radius for all the four indicators. Water quality of the samples in these clusters is more likely to change along with the changing environment.

Since researchers cannot monitor the water quality data ourselves, the data quality is not very good. Due to the shortage of big enough data, there are also some limitations on the use of data. All these shortcomings can be overcome by further research. More data can be obtained by the government or other organizations.

Table 2
Interval-valued data of sampling sites

Sample sites	pH	DO	COD_{Mn}	NH_3–N
1	[6.53,8.10]	[5.18,11.50]	[3.00,3.30]	[0.10,0.63]
2	[7.26,8.17]	[4.88,12.10]	[3.00,6.20]	[0.18,0.94]
3	[7.44,8.79]	[5,11.7.00]	[2.10,5.10]	[0.10,0.93]
4	[7.04,8.69]	[3.63,12.70]	[2.70,4.10]	[0.07,0.89]
5	[7.42,8.51]	[4.71,11.90]	[2.50,4.50]	[0.06,1.06]
6	[6.97,8.50]	[4.99,14.20]	[2.10,6.20]	[0.29,1.26]
7	[6.57,8.67]	[4.17,12.70]	[4.10,8.00]	[0.17,1.40]
8	[7.24,8.21]	[6.36,12.40]	[1.70,3.30]	[0.13,0.74]
9	[7.25,8.56]	[2.94,13.00]	[2.20,14.30]	[0.22,2.42]
10	[7.37,8.80]	[0.74,13.40]	[2.80,7.10]	[0.12,2.32]
11	[7.36,8.90]	[2.29,12.40]	[3.90,12.30]	[0.10,2.41]
12	[7.70,9.21]	[0.94,11.60]	[5.80,22.40]	[0.19,14.70]
13	[7.70,8.93]	[5.85,11.20]	[3.50,7.40]	[0.09,1.97]
14	[6.48,8.50]	[3.25,17.00]	[3.70,10.80]	[0.08,0.72]
15	[7.62,8.97]	[0.79,9.50]	[7.10,19.70]	[0.21,20.40]
16	[7.52,8.85]	[3.83,17.90]	[2.10,11.30]	[0.04,1.17]
17	[7.57,9.16]	[0.13,21.20]	[3.90,6.80]	[0.02,0.47]
18	[7.49,8.70]	[3.05,12.60]	[2.40,9.80]	[0.04,0.62]
19	[6.97,8.45]	[3.14,12.10]	[4.60,110.80]	[0.26,5.07]
20	[6.94,8.06]	[5.15,10.80]	[3.30,5.80]	[0.09,0.94]
21	[7.12,8.77]	[5.07,8.91]	[2.70,5.20]	[0.26,0.62]
22	[7.23,8.68]	[2.94,17.60]	[3.40,30.50]	[0.09,1.82]
23	[7.28,8.53]	[5.12,16.40]	[2.30,6.50]	[0.13,1.19]
24	[7.19,8.60]	[3.4,15.90]	[3.50,13.20]	[0.12,2.63]
25	[6.16,8.78]	[3.43,15.30]	[2.60,9.00]	[0.07,1.72]
26	[6.71,8.89]	[5.58,17.10]	[2.30,8.60]	[0.11,2.10]

Table 3
CRI from four to seven clusters

CRI	4 clusters	5 clusters	6 clusters	7 clusters
3 clusters	0.3473			
4 clusters		0.7989		
5 clusters			0.8505	
6 clusters				0.8442

Table 4
Mean of interval-valued data for each cluster

	pH	DO	COD_{Mn}	NH_3–N
1	[6.98,8.48]	[4.66,12.78]	[2.98,6.89]	[0.14,1.08]
2	[7.27,8.64]	[3.01,12.75]	[2.55,7.50]	[0.12,1.67]
3	[7.03,8.95]	[1.57,16.30]	[3.47,9.37]	[0.06,1.53]
4	[7.70,8.93]	[5.85,11.20]	[3.50,7.40]	[0.09,1.97]
5	[7.24,8.69]	[3.57,13.73]	[3.59,13.69]	[0.14,5.38]
6	[7.25,8.65]	[3.49,15.00]	[3.35,61.05]	[0.15,3.12]

Table 5
Radius of each cluster

	pH	DO	COD_{Mn}	NH_3–N
1	1.50	8.11	3.91	0.93
2	1.37	9.75	4.95	1.56
3	1.92	14.73	5.90	1.47
4	1.23	5.35	3.90	1.88
5	1.45	10.16	10.10	5.24
6	1.41	11.52	57.70	2.97

This method is different from the traditional water quality methods, so it is difficult to make a comparison. In the further study, it is recommended to evaluate the results from this method.

4. Conclusion

Based on the traditional data analysis method, interval-value data analysis is introduced to reduce the dimension of the water quality data and reduce the losses of information. The standardization is similar as that of point data. Then, interval data cluster analysis is used to assess water quality in Huaihe River. Euclidean–Hausdorff distance is employed to classify the samples into several clusters. In order to determine how many clusters there should be, CRI is employed. CRI is an indicator to measure the difference between two clusters. At the end, 26 sampling sites are classified into six clusters. Indicator concentrations of sampling sites in the same cluster fluctuate in a similar manner and the pollution of them has little difference.

This is just initial research for interval-valued data to assess water quality. The results show that this method is useful for water quality clustering. When the data are very big, the advantages of this method will be more obvious.

However, without comparison, it is difficult to find out whether this method is applicable or not. In the future, the research aims to employ other similar methods to check out the effect of the method. Also, in each element of interval data, data are collected from only 51 points. If more samples can be collected, the effect of interval data analysis will be more significant.

References

[1] Z. Zou, Y. Yun, J. Sun, Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment, J. Environ. Sci., 18 (2006) 1020–1023.

[2] H. Boyacioglu, H. Boyacioglu, Detection of seasonal variations in surface water quality using discriminant analysis, Environ. Monit. Assess., 162 (2010) 15–20.

[3] S. Shrestha, F. Kazama, T. Nakamura, Use of principal component analysis, factor analysis and discriminant analysis to evaluate spatial and temporal variations in water quality of the Mekong River, J. Hydroinf., 10 (2008) 43–56.

[4] X. Xin, W. Lu, L. Gong, Discriminant analysis method application in water quality assessment, Environ. Sci. Technol., 31 (2008) 113–115.

[5] W. Lu, J. Li, F. Yu, G. Yu, L. Liu, Application of step wise discriminant analytical method in screening factor in the water quality evaluation, J. Jilin Univ., 39 (2009) 126–30.

[6] A. Papaioannou, A. Mavridou, C. Hadjichristodoulou, P. Papastergiou, O. Pappa, E. Dovriki, I. Rigas, Application of multivariate statistical methods for groundwater physicochemical and biological quality assessment in the context of public health, Environ. Monit. Assess., 170 (2010) 87–97.

[7] S. Kamble, R. Vijay, Assessment of water quality using cluster analysis in coastal region of Mumbai, India, Environ. Monit. Assess., 178 (2011) 321–332.

[8] L. Zadeh, Fuzzy sets, Inf. Control, 8 (1965) 338–353.

[9] X. Wang, Z. Zou, H. Zou, Water quality evaluation of Haihe River with fuzzy similarity measure methods, J. Environ. Sci., 25 (2013) 2041–2046.

[10] S.C. Jiang, S.B. Ge, X. Wu, Y.M. Yang, J.T. Chen, W.X. Peng, Treating *n*-butane by activated carbon and metal oxides, Toxicol. Environ. Chem., 99 (2017) 753–759.

[11] Q. Wang, Z. Zou, Application of BP neural network in water quality assessment for Miyun reservoir recharged with reclaimed water, Acta Sci. Circumstantiae, 34 (2014) 2413–2416.

[12] T. He, J. Li, H. Huang, Water quality evaluation of RBF neural network based on optimized parameter of genetic algorithm, Comput. Eng., 37 (2011) 13–15.

[13] Y. Hu, H. Wang, A new data mining method based on huge data and its application, J. Beijng Univ. Aeronaut. Astronaut., 17 (2004) 40–44.

[14] H.H. Bock, E. Diday, Analysis of Symbolic Data, Springer-Verlag, New York, NY, 2000.

[15] W. Li, J. Guo, Methodology and application of regression analysis of interval-type symbolic data, J. Manage. Sci. China, 33 (2010) 38–43.

[16] A. Jain, M. Murty, P. Flynn, Data clustering: a review, ACM Comput. Surv., 31 (1999) 264–323.

[17] A. Gordon, Classification, Chapman and Hall, Boca Raton, FL, 1999.

[18] A. Sharma, R. Ganguly, A.K. Gupta, Impact assessment of leachate pollution potential on groundwater: an indexing method, J. Environ. Eng., 146 (2020) 116–131.

[19] R. Rana, R. Ganguly, A.K. Gupta, Indexing method for assessment of pollution potential of leachate from non-engineered landfill sites and its effect on ground water quality, Environ. Monit. Assess., 190 (2018) 1–23.

[20] A. Gibrilla, E.K.P. Bam, D. Adomako, S. Ganyaglo, S. Osae, T.T. Akiti, S. Kebede, E. Achoribo, E. Ahialey, G. Ayanu, E.K. Agyeman, Application of water quality index (WQI) and multivariate analysis for groundwater quality assessment of the Birimian and cape Coast Granitoid Complex: Densu River Basin of Ghana, Water Qual. Exposure Health, 3 (2011) 63–78.

[21] C. Güler, G.D. Thyne, J.E. McCray, K.A. Turner, Evaluation of graphical and multivariate statistical methods for classification of water chemistry data, Hydrogeol. J., 10 (2002) 455–474.

[22] S. Manikandan, S. Chidambaram, A.L. Ramanathan, M.V. Prasanna, U. Karmegam, C. Singaraja, P. Paramaguru, I. Jainab, A study on the high fluoride concentration in the magnesium-rich waters of hard rock aquifer in Krishnagiri district, Tamilnadu, India, Arabian J. Geosci., 7 (2014), 273–285.

[23] H. Liu, Z. Liu, Recycling utilization patterns of coal mining waste in China, Resour. Conserv. Recycl., 54 (2010) 1331–1340.

[24] L. Zhang, Y. Jia, L. Zhang, H. He, C. Yang, M. Luo, L. Miao, Preparation of soybean oil factory sludge catalyst by plasma and the kinetics of selective catalytic oxidation denitrification reaction, J. Cleaner Prod., 217 (2019) 317–323.

[25] H. Wang, H. Zhong, G. Bo, Existing forms and changes of nitrogen inside of horizontal subsurface constructed wetlands, Environ. Sci. Pollut. Res., 25 (2018) 771–781.

[26] E. Diday, F. Brito, Symbolic Cluster Analysis, O. Opitz, Eds., Conceptual and Numerical Analysis of Data, Springer-Verlag, Heidelberg, 1989, pp. 45–84.

[27] F. Carvalho, M. Csernel, Y. Lechevallier, Clustering constrained symbolic data, Pattern Recognit. Lett., 30 (2009) 1037–1045.

[28] F. Carvalho, P. Brito, H.H. Bock, Dynamic clustering for interval data based on l2 distance, Comput. Stat., 21 (2006) 231–250.

[29] C. Tenorio, F. Carvalho, J. Pimentel, A Partitioning Fuzzy Clustering Algorithm for Symbolic Interval Data Based on Adaptive Mahalanobis Distances, Proceedings of 7th International Conference on Hybrid Intelligent Systems, Kaiserlautern, 2007, pp. 174–179.

[30] F. Carvalho, C. Tenorio, Fuzzy k-means clustering algorithms for interval-valued data based on adaptive quadratic distances, Fuzzy Sets Syst., 161 (2010) 2978–2999.

[31] F. Carvalhoa, Y. Lechevallierb, Partitional clustering algorithms for symbolic interval data based on single adaptive distances, Pattern Recognit., 42 (2009) 1223–1236.

[32] A. Irpino, R. Verde, Dynamic clustering of interval data using a wasserstein-based distance, Pattern Recognit., 29 (2008) 1648–1658.

[33] S. Ren, J. Lv, Genetic algorithm-based kernel function FCM clustering algorithm for interval numbers, J. Syst. Eng., 23 (2008) 611–616.
[34] C. Yu, Z. Fan, A FCM cluster algorithm for multiple attribute information with interval numbers, Oper. Res. Manage. Sci., 13 (2010) 12–16.
[35] W. Li, H. Dai, J. Guo, Hierarchical clustering of generally distributed interval symbolic data, J. Appl. Stat. Manage., 32 (2013) 1071–1078.
[36] A.M. Danby, M.D. Lundin, B. Subramaniam, Valorization of grass lignins: swift and selective recovery of pendant aromatic groups with ozone, ACS Sustainable Chem. Eng., 6 (2018) 71–76.