Evaluation of water quality in mangrove ecosystem for the Beilun Gulf and Zhenzhu Bay: A Complex Approach

Minghui Ou$^{1,2}$, Shan Liang$^{*1,2}$, Qingyu Xiong$^{1,3}$ and Ru Zhang$^{1,2}$

1 Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, China
2 College of Automation, Chongqing University, Chongqing, China
3 College of Software Engineering, Chongqing University, Chongqing China
E-mail: omhcumt_1992@163.com, lightsun@cqu.edu.cn, cquxqy@163.com

Abstract. Water quality evaluation is vital for water monitoring and management in mangrove ecosystem of Beilun gulf and Zhenzhu bay in Guangxi of China. Over the past few years, with the expansion of urbanization and economic development, the coastal water quality in mangrove ecosystem is becoming worse. This literature applied the Fuzzy C-Means (FCM) and principal component analysis (PCA) techniques to evaluate the water quality in the study area and to evaluate the overall rating of stations with respect to the 14 water quality parameters. The PCA analysis is based on the precise clustering of FCM analysis. More reasonable evaluation of water quality by PCA was obtained. Our numerical analysis has verified the result of the current paper is correct. In addition, those results can provide some selectable methods to improve station number for interrelated management department with the limit funds and resources.

1. Introduction
Mangroves have global distribution within coastal tropical and subtropical regions, which are particularly sensitive to changes in the surrounding sea environment and provide many ecosystem services. It is well known that the mangrove ecosystem possesses many ecological functions [1], [2]. What’s more, mangroves and other coastal vegetation play a vital role in protecting the safety of human inhabitation and preventing wealth from destroying by tsunami and typhoon. The knowledge that mangroves may protect coastal communities from coastal hazards (coastal erosion, typhoon, tidal bores and salt spray, tsunami, etc.) is well known in tropical and subtropical coastal ecology and increasingly by coastal managers [3]. Accordingly, much attention have been paid to discuss the ecological effects of mangrove ecosystem from researchers in recent years, such as attenuate wave energy [4], [5], provide optimal breeding, feeding and nursery grounds for many ecologically and the valuable sources of fuel wood, fodder, timber, tannin and other natural products for local people [6], [7].

However, with the socioeconomic development and expansion of urbanization, mangrove ecosystem is facing a serious threat. Mangrove forest is extremely sensitive to the health of the growing environment, such as the soil, water and microbial environment. Indeed, the water quality has a great effect on the growth of mangrove forests and is prone to changed by the external environment. Many literatures have reported on the water quality parameter and monitoring station by some statistical methods. For example, the principal component analysis (PCA) and principal factor analysis (PFA) techniques have been used to evaluate the effectiveness of the surface water quality-monitoring network in a river and to analysis the water quality for the Nakdong river watershed [8], [9]. PCA
seeks to establish combinations of variables capable of describing the principal tendencies observed while studying a given matrix [10].

According to the long coastline for Beilun Gulf and Zhenzhu Bay in Guangxi of China, the monitoring devices have been distributed along with the long coastline. It is not hard to know that the effect factors of water quality between monitors which have long distance are different. To find a more precise method reflects the water quality is vital importance obviously.

In this paper, the monitoring station’s location information is known. And we can’t partition those stations into small clustering by eyes or experience exactly. The location information (Longitude and Latitude) were clustered into smaller groups using Fuzzy C-Means (FCM) algorithm. And then the monitoring data of stations have strong correlation in each group. PCA is a multivariable statistical technique which is used to identify important components that explain most of the variances of a complex system. This technique was used to process the data. Analysis the water quality of each clustering group, and the results will helpful to management the mangrove environment.

The contributions of this work are four-fold: 1) Get more precise clustering of location information by FCM algorithm. 2) Clustering process to enhance the correlation of monitoring data. 3) More reasonable evaluation of water quality for each partition group by PCA. 4) Provide some selectable method to reduce station number for interrelated management department with the limit funds and resources.

2. Method
In this section two methods of data process will briefly be introduced. The water quality of mangrove forest has many influencing factors (human activities, environment pollution and other uncertainties) along with the long coastline. Finding strong correlation monitoring data is a vital issue. This is very important, especially, to get the mainly principal component of water parameter in mangrove forest by PCA. Indeed, the FCM is a partitioning-based clustering algorithm and the clustering group will be high-efficiency for PCA analysis. It has the property to group a set of objects into classes or clusters so that the objects within a cluster have high similarity. And the FCM has clear advantage over crisp and probabilistic clustering methods [11].

2.1. Clustering
FCM is a fuzzy clustering algorithm which assigns memberships to each data corresponding to all clusters based on the distances of the data and the cluster centres [12]. Due to the characteristics of location information (Longitude and Latitude), it is not accurate to group location information into exclusive clusters. Thus, FCM is very suitable to solve the location information clustering problem as it use a fuzzy membership which assigns a number of membership for every cluster. The accurate description of FCM theory can be found in [12], [13], [14]. The basic FCM algorithm aims to minimize the objective function (1) iteratively.

$$J_o(U,G) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \| x_i - g_j \|$$

where $X = (x_1, x_2, \cdots, x_N)$ denotes the data samples set and $G = (g_1, g_2, \cdots, g_C)$ denotes the clustering groups, and $U = (u_{ij})_{N \times C}$ is a fuzzy partition matrix composed of the membership of each feature vector $x_i$ in each cluster $j$ and $u_{ij}$ should satisfy $\sum_{j=1}^{C} u_{ij} = 1$. The $m = 1, 2, \cdots, \infty$ is an index that controls the fuzziness of the resulting partition. $\| \|$ denotes the distance between data $x_i$ and the cluster centre $g_j$. If $C$ is an integer, $2 \leq C < N$, a conventional $C$-partition of $X$ is a $C$-tuple $(X_1, X_2, \cdots, X_C)$ of subsets of $X$ that satisfies three conditions:

$$X_i \neq \Phi \quad 1 \leq i \leq C; \quad (2)$$

$$X_i \cap X_j = \Phi \quad i \neq j; \quad (3)$$

$$X_i \neq \Phi \quad 1 \leq i \leq C; \quad (2)$$

$$X_i \cap X_j = \Phi \quad i \neq j; \quad (3)$$
\[ \bigcup_{j=1}^{C} X_j = X \]  

(4)

In these equations, \( \emptyset \) stands for the empty set, and \( (\cap, \cup) \) are intersection and union, respectively.

The objective function \( J_p \) is minimized when the sampled data set are assigned to closer to the cluster centres, and vice versa. The memberships and the cluster centres are computed by:

\[ u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{1}{d_{ik}} \right)^{\frac{1}{m}}} \]

(5)

\[ g_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \]

(6)

If \( m \) is set to 1, the memberships \( u_{ij} \) converge to 0 and 1, which transforms to hard clustering method.

2.2. Principle component analysis

PCA was first performed combined with the FCM to identify the potential for reducing the number of monitoring stations by analysing the sampled data in the monitoring area.

Mathematically, PCA is a multivariable statistical techniques used to identify components that explain most of the variances of a complex system by linear transportation. The procedures of PCA are shown as follows.

(i) Standardization of the measurements to ensure that they all have equal weights in the analysis,
(ii) Calculate the covariance matrix \( C \),
(iii) Find the eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_n \) and the corresponding eigenvectors \( \alpha_1, \alpha_2, \ldots, \alpha_n \) and extract the principal component,
(iv) Find the principal component comprehensive evaluation function.

The sampled data have strong correlation in every cluster. We should select high standard for the factor correlation coefficient. And the detailed procedure of the experiment is illustrated in Figure 1.

![Figure 1. Block diagram of the experiment procedure.](image)

3. Result and discussion

3.1. Location information clustering
In practice, the monitoring devices were installed along with the coastline and the estuaries to monitor water pollution. We got the location information of each monitoring station as shown in Table 1.

**Table 1.** Detail information (longitude and latitude) of monitoring station.

| Stations | Longitude   | Latitude   | Stations | Longitude   | Latitude   |
|----------|-------------|------------|----------|-------------|------------|
| 1        | 108.02194   | 21.54889   | 12       | 108.23889   | 21.58917   |
| 2        | 108.03500   | 21.54833   | 13       | 108.16139   | 21.52417   |
| 3        | 108.04333   | 21.54500   | 14       | 108.16111   | 21.52528   |
| 4        | 108.05528   | 21.53278   | 15       | 108.16056   | 21.52583   |
| 5        | 108.05972   | 21.52889   | 16       | 108.18361   | 21.61472   |
| 6        | 108.06639   | 21.52500   | 17       | 108.20917   | 21.60194   |
| 7        | 108.08611   | 21.53000   | 18       | 108.20454   | 21.59634   |
| 8        | 108.08833   | 21.53778   | 19       | 108.20500   | 21.58570   |
| 9        | 108.09056   | 21.54528   | 20       | 108.23000   | 21.61389   |
| 10       | 108.24806   | 21.59944   | 21       | 108.22944   | 21.61194   |
| 11       | 108.24583   | 21.59389   | 22       | 108.23000   | 21.60667   |

Those stations were clustered into 7 groups by using FCM algorithm. And the centre information also was obtained as shown in Table 2.

**Table 2.** The clustering result and centre coordinate

| Groups | I  | II | III | IV | V  | VI | VII |
|--------|----|----|-----|----|----|----|-----|
| Stations | 1,2,3 | 4,5,6 | 7,8,9 | 13,14 | 15,16,17,18,19 | 10,11,12 | 20,21,22 |
| Lon    | 108.003500 | 108.03500 | 108.04333 | 108.05528 | 108.05972 | 108.06639 | 108.08611 |
| Lat    | 21.54889   | 21.54833   | 21.54500   | 21.53278   | 21.52889   | 21.52500   | 21.53000   |
| Lon    | 108.08833   | 108.09056   | 108.08611   | 108.08611   | 108.08833   | 108.08833   | 108.24806   |
| Lat    | 21.53778   | 21.54528   | 21.53000   | 21.53000   | 21.53778   | 21.54528   | 21.59944   |
| Lon    | 108.24583   | 108.24583   | 108.24583   | 108.24583   | 108.24583   | 108.24583   | 108.24583   |
| Lat    | 21.59389   | 21.59389   | 21.59389   | 21.59389   | 21.59389   | 21.59389   | 21.59389   |

Where Lon and Lat are represent longitude and latitude, respectively. In order to get better observation for the clustering result. We have painted the seven groups on the map of study area by circle as shown in Figure 2.

**Figure 2.** Map information of study area and clustering result.

Figure 2 schematically illustrates the study area which is covered by mangrove forest and each cluster group was partitioned. The PCA analysis will be explored based on the partition.
3.2. **PCA analysis for each cluster**

Software IBM SPSS Statistic 19 and MATLAB 2014a are used to perform the analysis. However, the dissolved oxygen (DO) is a positive index. We need use the equation (7) to pre-processing the data of each DO.

\[
x_{DO} = \max x_{DO} - x_i
\]

One of our goals is to reduce the number of monitoring station of each cluster. The value Kaiser-Meyer-Olkin(KMO) is 0.743, which indicate to analysis the correlation and other properties are reasonable. The variance distribution in PCA for each cluster group is shown in Table 3. And the cumulative variance value is greater than 0.85(or 85%) was considered significant.

**Table 3.** The variance of the components.

| Components | Total  | Initial eigenvalue variance % | Cumulative value % |
|------------|--------|-------------------------------|--------------------|
| 1          | 7.2525 | 51.802                        | 51.802             |
| 2          | 2.828  | 20.203                        | 72.005             |
| 3          | 1.492  | 10.656                        | 82.661             |
| 4          | 0.756  | 5.403                         | 88.064             |

From Table 3, we selected four components and their variance together accounted for about 88.064% of the total variance. Therefore, our discussions should focus only on the first four components.

The loading matrix of four components is shown in Table 4. Compared the loading absolute value of four components that we can know the fourteen water quality parameters information is reflected by the four components properly.

**Table 4.** The loading matrix of components about parameters

| Water quality parameters | 1     | 2     | 3     | 4     |
|--------------------------|-------|-------|-------|-------|
| PH                       | -0.289| -0.349| 0.013 | -0.053|
| Salinity                 | -0.359| -0.089| -0.003| 0.154 |
| DO                       | -0.118| 0.492 | 0.005 | 0.112 |
| COD                      | 0.188 | -0.403| 0.172 | -0.455|
| Phosphate                | 0.345 | 0.167 | 0.043 | -0.009|
| Inorganic-nitrogen       | 0.342 | 0.159 | 0.039 | 0.187 |
| Crude oil                | 0.231 | -0.318| 0.222 | -0.223|
| Chlorophyll-a            | 0.127 | -0.434| -0.173| 0.535 |
| Cu                       | -0.177| 0.034 | 0.633 | 0.018 |
| Pb                       | -0.185| 0.164 | 0.564 | 0.157 |
| As                       | -0.309| -0.222| -0.054| 0.311 |
| Nitrogen-total           | 0.361 | 0.103 | 0.034 | -0.002|
| Phosphorus-total         | 0.202 | -0.180| 0.401 | 0.253 |
| Ammonia-Nitrogen         | 0.316 | -0.048| 0.065 | 0.444 |

From the loading matrix in the analysis, the four components score, \( P_1, P_2, P_3, P_4 \), can be given as follows, respectively.
\[ P_1 = -0.289x_1 - 0.359x_2 - 0.118x_3 + 0.188x_4 + 0.354x_5 + 0.342x_6 + 0.231x_7 + 0.127x_8 \\
- 0.177x_9 - 0.185x_{10} - 0.309x_{11} + 0.361x_{12} + 0.202x_{13} + 0.316x_{14} \]  
\[ P_2 = -0.349x_1 - 0.089x_2 + 0.492x_3 - 0.403x_4 + 0.167x_5 + 0.159x_6 - 0.318x_7 - 0.434x_8 \\
+ 0.034x_9 + 0.164x_{10} - 0.222x_{11} + 0.103x_{12} - 0.180x_{13} - 0.048x_{14} \]  
\[ P_3 = 0.013x_1 - 0.003x_2 + 0.005x_3 + 0.172x_4 - 0.043x_5 + 0.039x_6 + 0.222x_7 - 0.173x_8 \\
+ 0.633x_9 + 0.564x_{10} - 0.054x_{11} + 0.034x_{12} + 0.401x_{13} + 0.065x_{14} \]  
\[ P_4 = -0.053x_1 + 0.154x_2 + 0.112x_3 - 0.455x_4 - 0.009x_5 + 0.187x_6 - 0.223x_7 + 0.535x_8 \\
+ 0.018x_9 + 0.157x_{10} + 0.311x_{11} - 0.002x_{12} + 0.253x_{13} + 0.444x_{14} \]  

And then, in order to understand the comprehensive level of the water quality of the monitoring stations in the study area. We use the four components corresponding variance value as the coefficient to construct a comprehensive evaluation function (12), where \( P_i (i = 1, 2, 3, 4) \) represents the four components score.

\[ S = 0.51802P_1 + 0.20203P_2 + 0.10656P_3 + 0.05403P_4 \]  

According to the comprehensive evaluation function, the water quality of 22 stations in study area in 2014 was ranked by the comprehensive score. And the relevant results for each partition also were shown in Table 5.

**Table 5. The rank of water quality and the relevant partitions**

| Partitions | Monitoring stations | Overall rating |
|------------|---------------------|---------------|
| IV         | 14                  | -1.317        |
| V          | 15                  | -1.211        |
| VI         | 12                  | -1.179        |
| VI         | 11                  | -0.863        |
| II         | 6                   | -0.819        |
| III        | 7                   | -0.727        |
| IV         | 13                  | -0.725        |
| V          | 18                  | -0.700        |
| VII        | 22                  | -0.684        |
| V          | 17                  | -0.653        |
| III        | 8                   | -0.580        |
| VI         | 10                  | -0.485        |
| VII        | 21                  | -0.332        |
| VII        | 20                  | -0.326        |
| II         | 5                   | -0.251        |
| V          | 19                  | 0.014         |
| V          | 16                  | 0.153         |
| II         | 4                   | 0.304         |
| III        | 9                   | 0.383         |
| I          | 3                   | 2.567         |
| I          | 1                   | 3.542         |
| I          | 2                   | 3.887         |

Based on the score of the main components and the comprehensive score, the water quality of mangrove forest can be quantitatively described. Since the water quality index is a low-quality index, the higher the score, the more serious water pollution. On the other hand, we can know each monitoring station belong to a specific partition. And the water quality of each station for every cluster can be obtained easily. For example, the second station is the worst and the third station is the best in the first cluster. As the same way, we can get which is best or worst for other cluster.

According to the overall rating in Table 5, some guidance can be obtained from these results to improve the coastal water quality of the mangrove forest when the funds and resources are limited. For
example, it is a prior task for us to monitor the water pollution of the station that have bad ranking in each cluster if the management with a limited budget.

4. Conclusion
This paper has demonstrate the application of the FCM and PCA techniques to evaluate the water quality monitoring stations located in Beilun Gluf and Zhenzhu Bay of Guangxi and to evaluate the overall rating of stations with respect to the 14 water quality parameters is reasonable. And analysing the strategy of mangrove forest water resource management based on clustering was considered significant. Using the FCM and PCA techniques to analysis water pollution in our study area have profound significance. Not only can save the government’s economic costs, but also can save the manpower resources. For another, the methods will be useful for other similarity research.

5. References
[1] Nagelkerken I, et al., The habitat function of mangroves for terrestrial and marine fauna: A review. Aquatic Botany, 2008. 89(2): p. 155-185.
[2] Zhang K, et al., The role of mangroves in attenuating storm surges. Estuarine, Coastal and Shelf Science, 2012. 102-103: p. 11-23.
[3] Wells S A R C, In the front line: shoreline protection and other ecosystem services from mangroves and coral reefs. 2006.
[4] Massel S R, K Furukawa and R M Brinkman, Surface wave propagation in mangrove forests. Fluid Dynamics Research, 1999. 24(4): p. 219-249.
[5] Blankespoor B, Dasgupta S and Lange G M, Mangroves as a protection from storm surges in a changing climate. Ambio, 2017. 46(4): p. 478-491.
[6] Rasolofo M V, Use of mangroves by traditional fishermen in Madagascar. Mangroves and Salt Marshes, 1997. 4(1): p. 243-253.
[7] Spaninks F A V B, Economic valuation of mangrove ecosystems: potential and limitations. 1997.
[8] OUYANG Y, Evaluation of river water quality monitoring stations by principal component analysis. Water Research, 2005. 39(12): p. 2621-2635.
[9] Jung K Y, et al., Evaluation of water quality for the Nakdong River watershed using multivariate analysis. Environmental Technology & Innovation, 2016. 5: p. 67-82.
[10] Bengrati‘ne K and T F Marhaba, Using principal component analysis to monitor spatial and temporal changes in water quality. Journal of Hazardous Materials, 2003. 100(1-3): p. 179-195.
[11] Sun H, S Wang and Q Jiang, FCM-Based Model Selection Algorithms for Determining the Number of Clusters. Pattern Recognition, 2004. 37(10): p. 2027-2037.
[12] Bezdek J C, R Ehrlich and W Full, FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984. 10(2-3): p. 191-203.
[13] Liu L, et al., A modified Fuzzy C-Means (FCM) Clustering algorithm and its application on carbonate fluid identification. Journal of Applied Geophysics, 2016. 129: p. 28-35.
[14] Jamshidi A, et al., Dynamic risk assessment of complex systems using FCM. International Journal of Production Research, 2017: p. 1-19.

Acknowledgements
The authors would like to thank experts for their helpful comments and suggestions, and this work is supported by the Science and Technology Major Special Project of Guangxi (GKAA17129002).