The Application of Cluster analysis and Inverse Distance-weighted Interpolation to Appraising the Water Quality of Three Forks Lake

Wang Kea, He Ping Cheng, Dong Yan and Chen Lin *a*

Chengdu Electromechanical College, Chengdu, Sichuan, China

Abstract. The purpose of this paper is to realize the sustainable development of Three Forks Lake and control pollution. It analyzes the water quality of Three Forks Lake in the following aspects: PH, NH3-N, total phosphorus, total nitrogen, permanganate index, transparency, TDS, DO, conductivity. The author did some researches on water environment of Three Forks Lake by means of Cluster analysis and Inverse Distance-weighted Interpolation. The study reveals that most water area of Three Forks Lake is polluted seriously because the water here is having a PH greater than 7 and a greater turbidity, and the water nearer to inland is harder, having more nitrogen, higher permanganate index, less dissolved oxygen. This paper not only theoretically studies the water quality of Three Forks Lake by cluster analysis, but also get a conclusion of pollution with the combination of the information from Figure 1, Figure 3 and Figure 4, which basically identical with the authority's data.

Keywords: Cluster analysis; Inverse Distance-weighted Interpolation; Three Forks Lake; water quality

1. Introduction.

General Plan for the First Stage at Three Forks Lake Cooperation Zone intends to develop an area with multi-functions, such as tourism services, business, leisure, culture, sport, real estate[4]. Water plays a leading role in environment. The correct and appropriate appraisal of water quality is important to allocate, use, protect, manage water resources because it provides science foundation for planning water and preventing water environment pollution. If we pay no attention to water environment, the whole tourist industry might walk into vicious circle, even collapse. In order to realize the sustainable

*E-mail address:* tongwang889@TOM.COM
development of Three Forks Lake and control pollution, we try to study the present pollution there with the aid of Cluster analysis and Inverse Distance-weighted Interpolation.

2. Natural Geographical Situation

Three Forks Lake is the second largest lake in Sichuan Province, located in Sancha Town, Jianyang City. It was listed in Famous Lakes in the World in 1993. The lake strandline is 240 km, from north to south, the lake extends 18 km, and from east to west, 7 km. The water area is 27 sq. km. and the widest part of the surface is about 6,000 km. Its water retention capacity, 0.227 billion cu m, is three times that of West Lake in Hangzhou. Three Forks Lake has 113 islands, more than 160 peninsulas, 3 sq km wetland. Owning to good ecological condition, there are lots of birds and rich natural resources [5].

3. Research Methods

3.1. Cluster analysis[3]

Cluster analysis is mainly used to study the classification of various objects or phenomena. The basic idea of this approach is to set each sample as a class of their own, and then define the distance between the samples (or a similar factor) and the distance between classes. Select from the smallest pair, merge into a new class, and calculate the new class and the distance between the other classes, and then merging the closest two. This will reduce the number of the classes, one at a time, until all of the samples are distributed into one same class. According to the definition of between-class distance method, which is divided into the shortest distance, the distance method, the middle distance, center of gravity method, group average method, variable group average method, Ward from the sum of squares method, two-stage density estimation and so on. Cluster analysis of various systems comparative studies have shown that the majority of cases with good comprehensive properties and the Ward method and squared deviation class average, while the shortest distance method is the worst. Therefore, the author selected Ward sum of squares and the method of water pollution of Lake Sancha cluster analysis.

3.2. Square sum of deviations

Square sum of deviations is the total sum of squares. It is used to represent the distance between the two samples, and gradually cluster. Assuming a step class to class \( C_K \) and class \( C_L \) into the next level class \( C_M \). The \( B_{KL} \) is defined as the combined result of the incremental square sum of deviations when a step is class \( C_K \) and class \( C_L \) into the next level class \( C_M \), \( B_{KL}=W_M-W_K-W_L \), in which \( W_M \) is square sum of deviations of class \( C_M \), \( W_L \) is square sum of deviations of class \( C_L \). To represent the distance between two observation points potatoes with \( d(x,y) \) and represent \( D_{KL} \) the distance between class \( C_K \) and \( C_L \). The Formula of Ward square sum of deviations is:

\[
D_{KL} = B_{KL} = \left\| \bar{X}_K - \bar{X}_L \right\|^2 / (1/N_K + 1/N_L) \quad (1)
\]

Where, \( N_L \) is the number of observations in class \( CL \); \( \bar{X}_L \) is the mean vector of the class \( CL \), when the observation distance is

\[
d(x,y) = \left\| x - y \right\|^2 / 2 ,
\]

the recursive formula is:

\[
D_{JM} = (N_KD_{JK} + N_LD_{JL}) / N_M - N_KN_LD_{KL} / N_M^2 \quad (2)
\]
3.3. Interpolation

Interpolation is an important method of discrete function approximation. It can estimate the approximation value in other points through value status of the function in certain points. Common interpolation methods include Lagrange interpolation (lagrange interpolation), linear interpolation, Hermite and cubic spline interpolation, inverse distance weighted average interpolation.

Inverse Distance-weighted Interpolation is a method of partial interpolation, and its assumption is that the point of unknown value is influenced more by nearer observation points than by closer observation points. Since the changes in water quality keep are continuous, and the water qualities are more familiar in closer observation points, the weighted average of inverse distance interpolation is used in this article.

3.4. Inverse Distance-weighted Interpolation[1,2]

Calculating a grid node, the data point to give a specific weight of the specified node from the second side to the observation point is given of the node is proportional to inverse distance. When calculating a grid node, the weight is assigned to a fraction, the sum of weight equal to 1.0 of ownership. When an observation point with a grid node overlap, the actual observation point is given a weight of 1.0, all other observation points is given a weight of almost 0.0. In other words, the node is assigned a value consistent with the observation point. The formula is:

\[
f(x, y) = \begin{cases} 
  f_k & r_k = 0 \\
  \sum_{i=1}^N \frac{f_i}{r_i^2} & r_k \neq 0
\end{cases}
\]

Where, f(x,y) is the point for the coordinates in the interpolated value of (x,y), f_k is the distance between the observation points (x,y) for the r_k value.

4. Appraising the Water Quality of Three Forks Lake

4.1. Appraisal Standard and Parameter Selection

The major pollution source comes from cage culture of fish, household garbage, and wastes left by tourists. [4]According to the relative standards in <Quality Standards for Surface Water Environment> (GB3838-2002), and the water pollution of Three Forks Lake, we select 9 items as parameters, which are
PH, NH3-N, total phosphorus, total nitrogen, permanganate index, transparency, TDS, DO, conductivity.

4.2. Observation Point Selection.

The number of sampling points is limited due to the geographical problem in taking samples and the cost. We have studied the local water environment, set 16 representative points as observation points, and discretize Three Forks Lake geographically, using 182 points (Figure 1) to represent water quality of the lake. Now that the 16 observation points are known, we can calculate the value of 182 points resorting to the value of 16 points by means of Inverse Distance-weighted Interpolation.

| No. | Latitude  | Longitude | PH  | NH3-N | Total Phosphorus | Total Nitrogen | Permanganate Index | Transparency (m) | TDS | DO | Conductivity |
|-----|-----------|-----------|-----|-------|-----------------|---------------|--------------------|------------------|-----|----|-------------|
| 1   | 30.32     | 104.27    | 8.86| 0.27  | 0.09            | 0.87          | 4.90               | 1.03             | 115.8| 7.85| 257         |
| 2   | 30.31     | 104.27    | 8.86| 0.25  | 0.05            | 0.99          | 4.10               | 0.92             | 145.1| 7.60| 282         |
| 3   | 30.30     | 104.28    | 8.84| 0.26  | 0.05            | 0.87          | 4.27               | 0.82             | 124.0| 6.83| 283         |
| 4   | 30.30     | 104.29    | 8.77| 0.29  | 0.08            | 0.76          | 4.27               | 0.70             | 126.2| 6.47| 293         |
| 5   | 30.30     | 104.28    | 8.77| 0.23  | 0.04            | 0.58          | 3.87               | 0.88             | 127.0| 6.79| 293         |
We need to standardize original data with the following formula in that the dimension for testing is different, and the data differ greatly. During the process of standardizing all the data, we should unitize the type of data. Therefore we utilize various approaches to transform different types of data into the standard one. (The data process related to water quality of Three Forks Lake is not considered for water quality here is having PH greater than 7 and the data difference is not obvious.)

Table 2 Reference Figure for Contamination in Surface Water Environment

| Name | NH\textsubscript{3}-N | Total Phosphorus | Total Nitrogen | Permanganate Index | Transparency (m) | TDS | DO | Conductivity |
|------|----------------|------------------|----------------|-------------------|-----------------|-----|-----|-------------|
|      | Maxim \textsubscript{um} | Maxim \textsubscript{um} | Maxim \textsubscript{um} | Maxim \textsubscript{um} | Minim \textsubscript{um} | Maxim \textsubscript{um} | Minim \textsubscript{um} | Maxim \textsubscript{um} |
| Refere | 1 | 0.2 | 1 | 6 | 1 | 200 | 5 | 200 |
| Figure | $S_i$ |

And the transform formula of different style standard

And the transformation formula of different classification indexes is as follows (which $S_i$ is a reference value of surface water pollutants, specifically Table 2, for a measured pollutant concentration):

1) if the index is mini, then

$$x_i = \frac{S_i}{C_i}$$

2) if the index is max, then

$$x_i = \frac{C_i}{S_i}$$

Table 3 shows the standardized results of the examinations of pollutants.

Table 3  Actual Data of 16 Measured Sections in Three Forks Lake

| No. | Latitude | Longitude | PH(1) | NH\textsubscript{3}-N(2) | Total Phosphorus(3) | Total Nitrogen(4) | Permanganate Index(5) | Transparency(6) | TDS(7) | DO(8) |
|-----|----------|-----------|-------|--------------------------|-------------------|-------------------|----------------------|----------------|--------|-------|
| 6   | 30.29    | 104.28    | 8.88  | 0.26                     | 0.03              | 0.73              | 3.87                 | 0.81           | 119.0  | 7.58  | 275   |
| 7   | 30.29    | 104.29    | 8.78  | 0.29                     | 0.08              | 0.86              | 3.93                 | 0.84           | 121.7  | 6.69  | 285   |
| 8   | 30.28    | 104.29    | 8.80  | 0.28                     | 0.08              | 0.78              | 3.97                 | 0.73           | 127.2  | 6.99  | 297   |
| 9   | 30.25    | 104.28    | 8.87  | 0.45                     | 0.05              | 0.82              | 3.73                 | 0.78           | 129.6  | 7.51  | 291   |
| 10  | 30.23    | 104.27    | 8.88  | 0.41                     | 0.10              | 1.09              | 4.90                 | 0.63           | 146.4  | 9.57  | 338   |
| 11  | 30.30    | 104.25    | 8.84  | 0.27                     | 0.04              | 0.76              | 3.27                 | 0.91           | 128.4  | 6.59  | 293   |
| 12  | 30.30    | 104.23    | 8.71  | 0.55                     | 0.10              | 1.07              | 5.27                 | 0.63           | 206.0  | 6.59  | 471   |
| 13  | 30.26    | 104.27    | 8.80  | 0.35                     | 0.07              | 0.75              | 3.47                 | 0.75           | 124.6  | 7.42  | 287   |
| 14  | 30.32    | 104.26    | 8.86  | 0.33                     | 0.07              | 0.75              | 3.53                 | 0.82           | 117.7  | 9.00  | 258   |
| 15  | 30.26    | 104.26    | 8.86  | 0.43                     | 0.08              | 1.02              | 4.27                 | 0.61           | 122.7  | 4.63  | 297   |
| 16  | 30.30    | 104.26    | 8.79  | 0.30                     | 0.07              | 0.86              | 4.10                 | 1.11           | 121.2  | 7.05  | 280   |
Classify the data with the Ward method, and then use the toolbox of Classify in SPSS to analyze the standardized data of the 16 actual measured sectional ones

|    | X1  | X2  | X3  | X4  | X5  | X6  | X7  | X8  | X9  |         |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----------|
| 1  | 30.32 | 104.27 | 0.27 | 0.45 | 0.87 | 0.82 | 0.97 | 0.58 | 0.64 | 1.29 |
| 2  | 30.31 | 104.27 | 0.25 | 0.27 | 0.99 | 0.68 | 1.09 | 0.73 | 0.66 | 1.41 |
| 3  | 30.30 | 104.28 | 0.26 | 0.23 | 0.87 | 0.71 | 1.22 | 0.62 | 0.73 | 1.42 |
| 4  | 30.30 | 104.29 | 0.29 | 0.40 | 0.76 | 0.71 | 1.43 | 0.63 | 0.77 | 1.47 |
| 5  | 30.30 | 104.28 | 0.23 | 0.19 | 0.58 | 0.64 | 1.14 | 0.64 | 0.74 | 1.47 |
| 6  | 30.29 | 104.28 | 0.26 | 0.17 | 0.73 | 0.64 | 1.23 | 0.60 | 0.66 | 1.38 |
| 7  | 30.29 | 104.29 | 0.29 | 0.39 | 0.86 | 0.66 | 1.19 | 0.61 | 0.75 | 1.43 |
| 8  | 30.28 | 104.29 | 0.28 | 0.41 | 0.78 | 0.66 | 1.37 | 0.64 | 0.72 | 1.49 |
| 9  | 30.25 | 104.28 | 0.45 | 0.27 | 0.82 | 0.62 | 1.28 | 0.65 | 0.67 | 1.46 |
| 10 | 30.23 | 104.27 | 0.41 | 0.48 | 1.09 | 0.82 | 1.59 | 0.73 | 0.52 | 1.69 |
| 11 | 30.30 | 104.25 | 0.27 | 0.18 | 0.76 | 0.54 | 1.10 | 0.64 | 0.76 | 1.47 |
| 12 | 30.30 | 104.23 | 0.55 | 0.49 | 1.07 | 0.88 | 1.59 | 1.03 | 0.76 | 2.36 |
| 13 | 30.26 | 104.27 | 0.35 | 0.34 | 0.75 | 0.58 | 1.33 | 0.62 | 0.67 | 1.44 |
| 14 | 30.32 | 104.26 | 0.33 | 0.36 | 0.75 | 0.59 | 1.22 | 0.59 | 0.56 | 1.29 |
| 15 | 30.26 | 104.26 | 0.43 | 0.41 | 1.02 | 0.71 | 1.64 | 0.61 | 1.08 | 1.49 |
| 16 | 30.30 | 104.26 | 0.30 | 0.33 | 0.86 | 0.68 | 0.90 | 0.61 | 0.71 | 1.40 |

Fig. 3 final results of 16 sectional cluster; Fig. 4 cluster chart of pollution factors

1) Cluster according to cases

From Figure 3, Observation points 10, 12 and 15 belong to the same type and the rest belong to the other type. With the water distribution in Figure 1 being considered, the conclusion is drawn out that the three points 10, 12, 15 which belong to the river branches where the water is less and closer to the inland are more seriously polluted.

2) cluster through the pollution indices

The major indexes of pollution in Three Forks Lake are Index 1, Index 2, Index 5 and Index 8, which respectively refer to ammonia, total phosphorus, degree of transparency and electrical conductivity. Transparency and electrical conductivity are mainly influenced by oil and minerals. Without any factories nearby, Three Forks Lake is free from industrial discharge, therefore the major causes of pollution is from the cage breeding of fish and household garbages.

3) Interpolation
About the general situation of water quality of Three Forks Lake, the author means the method of weighted interpolation by inverse distance. From the observation points from 16 out of 182 points, the general situation of water pollution can be figured out as in Figure 2, in which point "*" is used to calculate the value of point "o". First, treat the latitude and longitude of point "o" as the coordinates of the interpolation point. Second, the value of feature parameters of the water quality can be drawn out with the help of formula (2.3). Similarly, with the method of hierarchical cluster of the analysis 182 water quality information, it’s conclude that the branches where observation points 10, 12, 15 are set is more seriously polluted than the other branches.

In summary, the environment in Three Forks lake has been gradually polluted, and the pollution of environment in branches where point 10, 12 and 15 were set has been in level IV[6]. Nowadays, the government has taken more measures to control the cage breeding in this area. Those people’s idea of protecting the environment has been improved and some equipments of breeding have gradually been removed.

Conclusion

Through a analysis of the following 9 parameters, which are PH, NH3-N, total phosphorus , total nitrogen, permanganate index, transparency, TDS, DO, and conductivity, the water quality in Three Forks Lake has been objectively estimated with the combination of the Cluster analysis and the method of Inverse Distance-weighted Interpolation. This provides foundations for the future plan and the continuous development. However, the latter method might be influenced by the local geographical situations and the number of observation points , and the conclusion can only partially reflect the real situation of the lake. Many discussions on this issue still need to be carried on

References

[1] Wu Chang Guan etc. Spatial Interpolation Methods and Temporal-Spatial Distribution of Precipitation in the Three Gorges Reservoir Area. Resources and Environment in the Yangtze Basin. 2010,7: 752-758
[2] WenRui Yang. Application of inverse distance weighted in terpolation method in contam inated site assessment, Chinese Journal of Applied Ecology, Sep, 2007, 18(9): 2013-2018
[3] Kang Na. Analysis of Regional Pollution Process and Application of System Cluster Method. Research of Environmental Sciences. Oct, 2009, 22(10): 1120-1127
[4] Environmental Assessment on General Plan for the First Stage at Three Forks Lake Cooperation Zone: http://www.sanchahu.com/xiangguan/2010_0404_55.html.
[5] Three Forks Lake in Jianyang City, http://www.hudong.com/wiki/%E7%AE%80%E9%98%B3%E4%B8%89%E5%B2%94%E6%B9%96.
[6] The Quality of Surface Water of January, 2011 in Sichuan Province, http://www.scemc.cn/structure/szyb/info?infid=155