Risk Identification of Water Pollution Sources in Water Source Areas of Middle Route of the South-to-North Water Diversion Project

Lian Sun, Meng Xu, Junxiang Jia, and Chunhui Li

Abstract—Water source areas are vulnerable to human's development. It is necessary to identify the risk of water pollution sources in water source areas with a large number of industrial and agricultural enterprises. An identification model based on entropy weight method and K-Means clustering analysis was proposed to identify the level of risk of water pollution sources in the water sources are qualitatively. Then the model was used in the case study Danjiangkou reservoir which is the water source of Middle Route of the South-to-North Water Diversion Project in China. The results show that there are nine industrial samples and two agricultural samples in the relatively high risk level which could engender high risk of water pollution to the Danjiangkou reservoir. In the end, suggestions are proposed to facilitate the management for water resources and environmental departments.

Index Terms—Clustering analysis, Danjiangkou reservoir, risk identification, water pollution.

I. INTRODUCTION

Water source areas are critical to human’s life, production and ecological system. With the rapid development of economy and society, the water source areas suffer the risk of anthropogenic water pollution. Therefore, it is essential to identify which are the higher or highest risk that could result in environmental risk. Risk identification denotes the process that we identify or classify the properties of potential or existing risk sources and factors by using certain methods. It is the first stage of risk management and is the basic of the next steps of risk management such as analysis and control of risk management [1].

Industry enterprises are usually the largest risk sources threatening the water sources districts. Peng, Song et al. [2] focused on the three aspects of emergency water pollution: risk sources, influence range and the sensitive receptor and developed a five-step method to assess the losses from potential chemical pollution accident in industrial parks. Zhao, Qin et al. [3] assessed the emergency health risk of emergency water pollution produced by the chemical company because of the emergency accident based on the quotients method and probability distribution of quotients qualitative. At the same time, the non-pointed sources pollution is becoming the important one besides the pointed pollution. Coupled the characteristics of NPS pollution, such as stochastic, broad, fuzzy and delayed effects [4], the research of risk identification are usually operated with geographic information system (GIS) and geostatistics method. Wang, Zuo et al. [5] established identification model of critical risk source areas for agricultural non-point source pollution with multifactor analysis method based on the indexes system of GIS. USLE(Universal Soil Loss Equation) is a general used model to identify the potential risk of NPS pollution. Hu, Wang et al. [6] modified the USLE model to identify the risk area with the help of GIS in a watershed. Combined with remote sense and GIS, Yang, Xu et al. [7] developed a model considering many factors contributing to the NPS pollution and used it identify the districts suffering from potential water pollution risk.

Analyzing with multi-factors is a fundamental and most frequently used method to determine the critical risk sources. The critic problem of multi-factor analysis is to establish the index system and determine the weight of assessment indexes [8]. Entropy is a function to describe the status of a system, and it could be used to reflect the order degree of a system. Entropy is a relatively ideal tool to gauge the weight of indexes in an index system and entropy weight method has the advantage in determining the index weight objectively [9]. K-means clustering analysis could achieve classification by assigning a set of objects into groups, called clusters, so that the objects in the same cluster are more similar. Different from other clustering methods, K-means clustering analysis has a merit of quickly and concise algorithm. Therefore, if a water quality manager use a method that integrates the entropy weight method and K-means clustering analysis, we would assess the risk more objectively and identify which risk sources should be paid more attention to supervise and manage.

Danjiangkou reservoir, lying in middle China, serves as the water source of Middle Route of the South-to-North Water Diversion Project which just began to transfer water to North China in the end of 2014. The aim of this project is expected to alleviate the severe water scarcity of agricultural production and sustain the increasing need of domestic department in the North China Plain. Consequently, it is of great importance to ensure the security of water quality in Danjiangkou reservoir. The issue to identify the risk of water pollution for water sources in Shiyuan is proposed necessarily.

In this paper, an identification model is developed to identify the risk of water pollution sources based on...
objective assessment method of index weight and clustering analysis method. Then the integrated method was applied to Danjiangkou reservoir as a case study to identify the risk level of water pollution sources and the distribution characteristics of these sources.

II. MATERIALS AND METHODS

A. Overview of the Studying Area

Shiyian district lies in the northeast of Hubei province in central China (Fig. 1). It is located in the north subtropics climate zone, with the annual average temperature of 13.7 °C and annual precipitation of 873 mm. Precipitation varies largely among different seasons, of which more than 80% occurs from May to August. The geomorphology in the basin is absolutely dominated by mountain with great variation of altitude and steep slope. The whole topography is tilted from west to east.

Danjiangkou reservoir, the water source of Yangtze River, the longest tributary of Yangtze River, Han River, Tianhe River et al. consist the river system of Shiyan. Danjiangkou reservoir, the water source of China's greatest water diversion project, was established in 1973 by closing the main watercourse of Hanjiang River in Danjiangkou district. The reservoir has the area of 1022 km² covering parts of Hubei and Henan province with the storage capacity of 29 billion m³. Recent year the water quality in reservoir is in the normal level of II~III yet still suffers the risk of water pollution especially the NPS pollution [5].

Shiyian district comprises of seven parts: Shiyian city, Danjiangkou city, Yunxian county, Yunxi county, Zhushan county, Zhuxi county and Fangxian county, with the whole territory of 2.4 × 10⁶ km² and the population of 3.5 × 10⁶. Economy and society develop quickly in the study area recent year. Since most of the area is in the upstream of the reservoir, the increasingly industrial development and expanding scale of livestock and poultry breeding in Shiyian pose more and more risk on the water sources.

B. Method

There are several kinds of perils of water pollution: traffic accident in land and water, sewage discharge of industry, and leakage of pipeline in the coast of water source areas. In recent years, the environmental department of Shiyian has taken series of strict management measures for Danjiangkou reservoir. The vessels pass the dam of Danjiangkou reservoir will be registered strictly, all the roads along the coast of the reservoir have established fences and it keeps certain distance between the fences and the canals of water diversion. With the altitude of dam was increased from 157m to 170m, about 345,000 people were emigrated out of the reservoir area, which reduces the risk to reservoir markedly. Therefore, industrial point source pollution and agricultural non-point source(NPS) pollution, threatening water quality. Since point and non-point resource pollution have different pollution mechanism, it’s essential to identify the risk level for these two respectively and select the high risk level to facilitate the water security management of a water source. As far as the study area in this paper, the water produced by livestock and poultry breeding is the absolutely dominant NPS pollution. Thus, the livestock and poultry breeding is chosen as the major target in agricultural department, ignoring others such as planting.

1) Standard pollution load

It is essential to sort out the major water pollution sources or main pollutants in a system which has more than one water pollution source and pollutant. As the amount and unit of discharge of different pollutants vary, usually it takes equal standard pollution load to standardize all the pollutants concerned.

\[
d_i = q_i / C_{oi} \times 10^6
\]

\[
r_i = d_i / \sum d_i \times 100\%
\]

where \(d_i\) is the equal standard pollution load of a pollutant(m³/a), \(q_i\) is the discharge amount of the sewage(t/a), \(C_{oi}\) is the assessment standard of \(i_{th}\) pollutant(mg/L). \(\sum d_i\) is the total load of equal standard pollution(m³/a), \(r_i\) is the load rate of \(i_{th}\) pollutant.

2) Entropy weight method to calculate weights

Entropy is a thermodynamic concept and it has been used in many domains including engineering technology and social economic. The principle of entropy weight method can be illustrated as: if the variation of an index value is larger, which means the amount of information of the index is richer and functions better in the comprehensive assessment as well as a larger entropy value and bigger weight [10]. Here is the operation steps of entropy weight method.

Assuming that there are s assessment indexes and t assessment samples, which constitutes a decision-making matrix A. Then we get the matrix \(D_i\) by standardizing matrix \(A\):

\[
D_i = \begin{bmatrix}
    d_{i1} & \cdots & d_{is} \\
    \vdots & \ddots & \vdots \\
    d_{i1} & \cdots & d_{its}
\end{bmatrix}
\]

\(d_{ij} \in [0,1](i = 1,2,\cdots, t; j = 1, 2, \cdots, s)\)

Then the entropy of \(f_{ij}(j=1,2,\cdots,s)\) index is defined as:

\[
E_j = -k \sum_{i=1}^{t} f_{ij} \ln f_{ij} \cdot j = 1,2,\cdots, s
\]
where \( f_{ij} = \frac{d_{ij}}{\sum_{i=1}^{t} d_{ij}} \), \( j = 1, 2, \ldots, s \), and if \( f_{ij} = 0 \), \( f_{ij}/n_{j} = 0 \), \( k = 1/n_{t} \).

And the entropy weight can be calculated:

\[
w_j = \frac{1 - E_j}{\sum_{j=1}^{s} (1 - E_j)} \cdot j = 1, 2, \ldots, s
\]

(5)

3) K-means clustering analysis

K-means clustering analysis was proposed by J. B. Macqueen to process the problem of clustering of data [11]. Now it is a widely used clustering analysis method because of its high efficiency and concise algorithm. This algorithm is a process that splits the data into K clusters according to a distant function after having determined the data set and center value of cluster K. In this paper the comprehensive assessment values \( Q_i \) are the process data which could be computed as:

\[
Q_i = \sum_{j=1}^{s} w_{ij} \times d_{ij}
\]

(6)

Here are the detail steps of the K-means clustering analysis.

Firstly we set \( k \) data as the initial cluster centers stochastically \((m_1, m_2, \ldots, m_k)\). The value set is relative to how many grades we plan to split.

Then each \( Q_i \) was divided into its nearest initial center, and a cluster center and its data are constituted as a cluster.

And based on the data divide into the cluster, each initial cluster center will recalculate the new cluster center \( m_i(7) \). This step is also called iteration.

\[
m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Q_{ij}, i = 1, 2, \ldots, k
\]

(7)

where \( N_i \) means there are \( N_i \) data in the cluster center \( m_i \), \( Q_{ij} \) means the \( j \)-th data in cluster center \( m_i \).

The iteration will be proceeded until it satisfies the termination condition that \( f(8) \), the sum of square error, is convergent. Therefore we can get the finally cluster centers \((m_1, m_2, \ldots, m_k)\).

\[
J = \sum_{i=1}^{k} \sum_{j=1}^{s} |Q_{ij} - m_{ij}|^2
\]

(8)

4) Livestock and poultry breeding analysis

The livestock and poultry in stock and output of stock were aggregated according to the statistical yearbook of Shiyan from 2009–2012 (Table I). Coupled with manure and urine generation coefficient of livestock and poultry (Table II), the total manure and urine could be estimated.

| TABLE II: MANURE AND URINE GENERATION COEFFICIENT OF LIVESTOCK AND POULTRY |
|-----------------------------------|-----------|----------|-----------|-----------|
|                                   | Cattle and buffaloes | Hog      | Sheep     | Poultry   |
|                                   | in stock    | output   | in stock  | output   |
| Manure                           | 25          | 10       | 15        | 10        |
| Urine                            | 10          | 3        | 0.5       | 0.5       |
| Breeding period                  | 365         | 365      | 365       | 90        |

III. RESULTS AND DISCUSSION

A. Industrial Risk

Building brick manufacturing, auto and auto parts manufacturing, metal processing and manufacturing, mining and dressing are the four main industries in Shiyan which account for the 80% companies.

Discharge quantity of sewage. Discharge concentration of industrial source. Discharge rate of sewage. Industry category and Distance to the reservoir were selected as the five risk identification indexes of industrial risk. COD, \( \text{NH}_4^-\text{N}, \text{petroleum}, \text{volatile phenol}, \text{cyanide}, \text{As}, \text{total Cr}, \text{Cr}^{6+}, \text{Pb}, \text{Cd}, \text{Hg} \), all this eleven indexes were selected as the second level index of Discharge concentration of industrial sources, according to the table of risk identification indexes of industrial sewage and discharge situation of water pollutant in Shiyan.

There are 292 industrial company samples, numbered as P1, P2, ..., P292, that could be the risk sources to the Danjiangkou reservoir in Shiyan. By standardizing pollution load and calculating the entropy of all the samples, we can get the weights of these indexes: 0.0393, 0.0601, 0.0605, 0.1091, 0.1001, 0.125, 0.150, 0.1135, 0.0943, 0.1062 and 0.1021 respectively. By weighted average of these eleven indexes, the weighted of the first level index Discharge concentration of industrial source was generated (the third column in Table III except its first value).

So operated as the index Discharge concentration of industrial source, the other four risk identification indexes of industrial risk can be calculated (Table III). Through weighted average, we can get comprehensive assessment value \( Q_i \) of each industrial sample. The detail weights and value of \( Q_i \) are displayed in Table III.

When the number of initial cluster center was set as 3, all the data were classified into their own cluster by using K-means clustering analysis. The result shows that there are nine pollution sources in the high cluster level 2 or 3, other 283 sources are in low cluster level 1. That means that most likely pollution sources are low risk of water pollution, only about 3% of the sources are high risk of water pollution.

According to the result of K-means clustering analysis, the point P40, P64, P128, P134, P138, P144, P145, P171 are in the relatively high risk level, and P89 is the only one point in the high risk level. On one hand, among these relatively high risk pollution sources, three of which (P134, P45, P171) are mining and dressing, two (P128, P138) are metal smelting, others are plastic manufacturing, thermal power, auto refitting, and plastic manufacturing (Table IV). On the other hand, if projecting them into the map, we can easily find out...
where the different risk level sources are located (Fig. 2). There are four points (P128, P138, P144, P145) lying in the coast of Tianhe River, a tributary of Danjiangkou reservoir; and the point P89, the highest risk level one, just lies in the coast of Danjiangkou reservoir. Additionally, five sources are in Yunxi lying in the upstream of Shiyan. These remind the water security managers that they should take more attention to the mining and dressing, metal smelting or the point P89 to reduce the risk of water pollution. At the same time, managers should focus on some tributaries, especially on the Tianhe River, and on certain district like Yunxi county. Other industries and districts should also be considered.

![Fig. 2. Clustering results of industrial pollution sources risk identification](image)

Note: several points overlap others for the limitation or the extension of the map.

**TABLE II: IDENTIFICATION RESULTS OF INDUSTRIAL RISK**

<table>
<thead>
<tr>
<th>Point</th>
<th>Discharge quantity of sewage</th>
<th>Discharge concentration of industrial source</th>
<th>Discharge rate of sewage</th>
<th>Industry category</th>
<th>Distance to the reservoir</th>
<th>Comprehensive assessment value</th>
<th>K-means clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>P89</td>
<td>0.0000</td>
<td>0.9752</td>
<td>0.0058</td>
<td>0.0060</td>
<td>0.0014</td>
<td>0.0357</td>
<td>2</td>
</tr>
<tr>
<td>P64</td>
<td>0.0908</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0060</td>
<td>0.0058</td>
<td>0.0240</td>
<td>2</td>
</tr>
<tr>
<td>P138</td>
<td>0.0100</td>
<td>0.0432</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.0005</td>
<td>0.0196</td>
<td>3</td>
</tr>
<tr>
<td>P131</td>
<td>0.0010</td>
<td>0.0713</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.0005</td>
<td>0.0196</td>
<td>3</td>
</tr>
<tr>
<td>P145</td>
<td>0.0002</td>
<td>0.1283</td>
<td>0.0013</td>
<td>0.0119</td>
<td>0.0023</td>
<td>0.0061</td>
<td>2</td>
</tr>
<tr>
<td>P171</td>
<td>0.0128</td>
<td>0.0599</td>
<td>0.0033</td>
<td>0.0060</td>
<td>0.0067</td>
<td>0.0321</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: Discharge concentration of industrial source refers to the discharge concentration of COD, NH₃-N, petroleum, volatile phenol, cyanide, As, total Cr, Cr⁶⁺, Pb, Cd, Hg; Discharge rate of sewage = Discharge quantity of sewage / Production quantity of sewage.

**TABLE IV: CLUSTERING RESULTS OF INDUSTRIAL RISK**

<table>
<thead>
<tr>
<th>Point</th>
<th>Cluster level</th>
<th>Location</th>
<th>Industrial sector</th>
<th>Receiving water body</th>
</tr>
</thead>
<tbody>
<tr>
<td>P40</td>
<td>2</td>
<td>Zhangwan district</td>
<td>Specialized chemical products manufacturing</td>
<td>Jianghe River</td>
</tr>
<tr>
<td>P64</td>
<td>2</td>
<td>Zhangwan district</td>
<td>Thermal power</td>
<td>Hongweihe River</td>
</tr>
<tr>
<td>P89</td>
<td>3</td>
<td>Yunxian</td>
<td>Auto refitting</td>
<td>Danjiangkou reservoir</td>
</tr>
<tr>
<td>P128</td>
<td>2</td>
<td>Yunxi</td>
<td>Puddling</td>
<td>Tianhe River</td>
</tr>
<tr>
<td>P134</td>
<td>2</td>
<td>Yunxi</td>
<td>Antimony ore processing</td>
<td>Xianhe River</td>
</tr>
<tr>
<td>P138</td>
<td>2</td>
<td>Yunxi</td>
<td>Antimony dressing</td>
<td>Tianhe River</td>
</tr>
<tr>
<td>P144</td>
<td>2</td>
<td>Yunxi</td>
<td>Plastic manufacturing</td>
<td>Tianhe River</td>
</tr>
<tr>
<td>P145</td>
<td>2</td>
<td>Yunxi</td>
<td>Lead and zinc ore dressing</td>
<td>Tianhe River</td>
</tr>
<tr>
<td>P171</td>
<td>2</td>
<td>Zhushan</td>
<td>Silver dressing</td>
<td>Duhe River</td>
</tr>
<tr>
<td>Others</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

There are 65 agriculture NPS pollution samples, numbered as L1, L2, ..., L65. Based on the data in Table V, standard pollution load and the entropy weight of these four indexes (Breeding scale, Producing amount of manure, Producing amount of urine, Discharge rate of pollutants) were conducted successively. After acquiring the comprehensive value Q, and setting the cluster center as 2, we could identify all the 65 data into their cluster level. The results show that there are only two points in the second risk level representing high risk of water pollution and other 63 points are in relatively low risk (Table VI).

There are 65 agriculture NPS pollution samples, numbered as L1, L2, ..., L65. Based on the data in Table V, standard pollution load and the entropy weight of these four indexes (Breeding scale, Producing amount of manure, Producing amount of urine, Discharge rate of pollutants) were conducted successively. After acquiring the comprehensive value Q, and setting the cluster center as 2, we could identify all the 65 data into their cluster level. The results show that there are only two points in the second risk level representing high risk of water pollution and other 63 points are in relatively low risk (Table VI).

Point L23 is an animal husbandry company located in the Yunxian county with the distance of 1.5 kilometers to the main stream of Hanjiang River. The company covers an area of 45 km², ranking to the company covering the largest area among all the 65 agricultural companies. L54 is a company whose business scope embraces poultry, cultivating of vegetable, fruit, flowers, et, al. It locates in the Langhe River, a tributary of Danjingkou reservoir. It covers an area of 5.86 km², much smaller than L23. However, the comprehensive assessment value of L54 is larger than L23, which shows the

**B. Agriculture Risk**

The amount of manure and urine of livestock and poultry in Shiyan district can be estimated by multiplying the data in Table I and Table II. As is shown in Table V, the total amount of manure produced by the five main kinds of livestock and poultry is 6645702 ton, and the urine is 4806564 ton.

NPS pollution produced by poultry breeding and livestock and is affected by a variety of factors, such as the amount of manure and urine of animal, the distance from the water area
risk level of L54 is higher. Both of these two companies are not far from the reservoir. Thus the managers should take care of their potential pollution threaten to Danjiangkou reservoir.

<table>
<thead>
<tr>
<th>Point</th>
<th>Breeding scale</th>
<th>Producing amount of manure</th>
<th>Producing amount of urine</th>
<th>Discharge rate of pollutants</th>
<th>Comprehensive assessment value Q,</th>
<th>Cluster level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>0.3081</td>
<td>0.3609</td>
<td>0.2805</td>
<td>0.0444</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L23</td>
<td>0.0537</td>
<td>0.2587</td>
<td>0.1118</td>
<td>0.0046</td>
<td>0.1416</td>
<td>2</td>
</tr>
<tr>
<td>L54</td>
<td>0.2686</td>
<td>0.1570</td>
<td>0.2067</td>
<td>0.0123</td>
<td>0.1989</td>
<td>2</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

Combined with standard pollution load, the K-means clustering analysis method was conducted to identify the water pollution risk for Shiyan district where it is the water resource of China’s Middle Route of the South-to-North Water Diversion Project. All the 192 industrial risk sources and 65 agricultural risk sources were evaluated respectively and the weights of which were calculated. Consequently we can identify which are in the higher ranks represented by the higher cluster center numbers. By analyzing the attribute of these higher points and checking their spatial distribution, we concluded some useful findings.

There are 9 industrial companies in the higher risk level 2 or 3. These companies usually belong to industrial sector of mining and dressing, metal smelting. Four companies of these nine are established in the coast of rivers and the only one company in the highest risk is lying to the Danjiangkou reservoir. The Yunxi county owes most of these high risk companies.

There are 2 agricultural companies in the higher risk level 2. One is an animal husbandry company lying the Yunxian county, and the other conducts producing of poultry, cultivating of vegetable, fruit, flowers, et. al. Both of them are covers larger area than industrial companies and have the risk of nonpoint water pollution. And these two companies are not situated far from the reservoir, which poses relatively high water pollution risk to the Danjiangkou reservoir.

These remind the water security managers that they should take more attention to certain industrial sector like the mining and dressing, metal smelting or certain point like P89 to reduce the risk of water pollution. At the same time, managers should focus on some tributaries, especially on the Tianhe River, and on certain district like Yunxi county. Regarding the agricultural companies, the managers in Shiyan should pay more attention to several companies like L23 and L54 since they are not far from the reservoir and have the highest risk level. Other industries and districts and the agriculture companies should also be considered.

ACKNOWLEDGMENT

This study was supported by the National Science and Technology Support Program (No. 2011BAC12B02).

REFERENCES


Lian Sun was born in Qimen, Anhui province, China. He obtained a bachelor degree of geography in Anhui Normal University. Now, he is a graduated student majoring in hydrology and water resources in the School of Environment, Beijing Normal University, Beijing, China.

Meng Xu was born in Taiyuan, Shanxi province, China. She received her master degree of environmental and resource management in Brandenburg University of Technology, Cottbus, in 2006 in Germany. Currently, she is a lecturer in School of Urban/Rural Planning and Management in Zhejiang University of Finance and Economics, majoring in water resource management. Since 2013, she has been the doctoral candidate in the School of Environment, Beijing Normal University, majoring in water resource management and allocation.

Junxiang Jia was born in Dezhou, Shandong province, China. She earned her degree of master of hydrology and water resources from Beijing Normal University, Beijing China. Now she is working in the Ministry of Water Resources of P. R. China.

Chunhui Li was born in Linquan, Anhui province, China. He received his Ph.D degree from Beijing Normal University, China, and did research as a post-doctor in Regina University, Canada. Now he is an associated professor doing research in water resources assessment and planning and environmental influence assessment in the School of Environment, Beijing Normal University.