Scholar Publishing Group
Water Pollution Prevention and Control Project
https://doi.org/10.38007/WPPCP.2021.020303
ISSN 2790-2730 Vol. 2, Issue 3: 22-31

Construction of Water Pollution Early Warning System Based on Clustering Algorithm and Machine Learning

Janeth Arias*

Universitas Sebelas Maret, Surakarta 57126, Indonesia

**corresponding author*

Keywords: Water Pollution, Monitoring and Early Warning System, Clustering Algorithm, Absolute Error Average, Mean Squared Error

Abstract: With the rapid development of industry and towns, environmental pollution accidents occur frequently. It is of great significance to establish a sound environmental monitoring and early warning system to ensure the safety of water environment and people's water use. On this basis, this paper deeply discussed the early warning of water pollution, and proposed a water pollution early warning system based on clustering algorithm. The research showed that when the water pollution early warning system was used to monitor and warn the harmful substances such as formaldehyde, cyanide and Chemical Oxygen Demand (COD), the average absolute error of the water pollution early warning system when monitoring the four harmful substances was about 0.67, and the mean square error of the water pollution early warning system when monitoring the four harmful substances was about 0.66. From the perspective of the mean absolute error and mean square error, it could be seen that the water pollution early warning system based on clustering algorithm had good detection accuracy. In addition, in the evaluation study of the water pollution early warning system, this paper reached the conclusion that most of the monitoring and early warning technicians were relatively approved of the water pollution early warning system.

1. Introduction

In recent years, human activities and changes in the natural environment have caused serious damage to the water resources and environment. At the same time, various human factors and accidental water quality emergencies occur frequently, thus seriously threatening the safety of urban water supply system. Cluster analysis is an important technology in the field of data mining, which can be well applied to data mining of digital types and the implementation of algorithms. By facing

this situation, this paper proposed a water pollution early warning system based on clustering algorithm, thus hoping to provide valuable reference for relevant research.

Many scholars have studied water pollution and water quality detection. Hossain SM Zakir briefly introduced the biological monitoring technology and described its application in monitoring pesticide content and heavy metal content [1]. In order to predict the two water quality variables of Little Prespa Lake in Greece, Barzegar Rahim developed a short-term memory model, a convolution neural network model and a mixed model of the former two for water quality monitoring [2]. Echavarria-Caballero Carolina investigated and studied the water quality evolution in Surest Park and concluded that the water quality in Surest Park improved [3]. Li Z. H. O. U. studied the impact of river head system on water pollution, and concluded that the implementation of long system could reduce the negative impact of livestock manure production on surface water quality [4]. Khatri Punit developed a new water quality monitoring system for the limitation of high cost of water quality monitoring and proved its excellent monitoring performance through experiments [5]. Alam Arif U developed a monitoring system which could simultaneously monitor acid-base values, free chlorine, heavy metals and other substances in the water environment [6].

Water pollution and water quality detection are of great significance, and the following scholars have also participated in their research. Wu Jianhua studied the water quality and pollution of groundwater, and explored the main influencing factors of groundwater pollution after analyzing the water quality parameters [7]. Chen Zeng studied the noise removal of water quality data and proposed an adaptive noise removal method, which provided a reference for the relevant research on noise removal of water quality data [8]. Zhu Weiyu developed a method for surface water quality detection based on regression model, which could effectively detect water quality anomalies [9]. Pujar Prasad M. used the Internet of Things method to collect water quality data, and carried out water quality practice analysis of the Krishna River [10]. In order to better identify groundwater pollution, Xia Xuemin proposed a super parameter adjustment robust method to optimize the performance of genetic algorithm, and tested the effectiveness of this method through experiments [11]. In order to protect the health of water environment, this paper studied the construction of water pollution early warning system based on clustering algorithm and machine learning.

In order to ensure the health of the water environment and the safety of people's living and production water, this paper studied the early warning of water pollution, and proposed a system for early warning of water pollution based on clustering algorithm and machine learning. It also analyzed the overall structure and functions of the system, and discussed the control methods of water quality monitoring. This paper also studied the monitoring and early warning effect of the water pollution early warning system. Compared with other studies, the water pollution early warning system proposed in this paper used clustering algorithm as the data mining method, and the monitoring and early warning results also proved that the system had certain effectiveness.

2. Overall Structure of Water Pollution Early Warning System

The water source pollution early warning system includes three modules: water source sampling, data transmission and early warning, as shown in Figure 1.

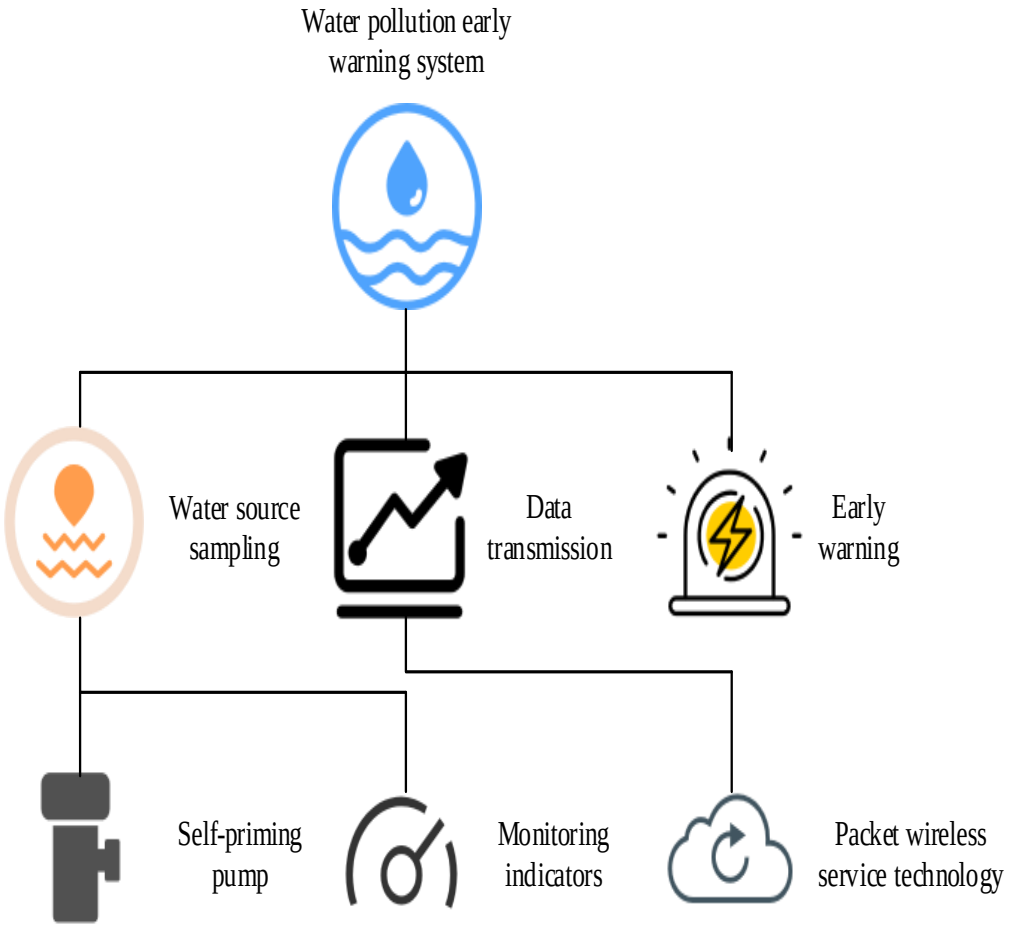


Figure 1. The overall architecture of the water pollution early warning system

The sampling module uses two self-priming pumps. One is used while the other is standby. A pressure sensor is installed on the outlet pipe of each pump. In routine monitoring, the system controls a water pump to start and take water samples. If the running water pump fails, the pressure sensor would detect the water level drop in the pipeline, and the system would stop the started water pump and convert it to other water pumps. In the selection of monitoring indicators, multiple pollution sources should be considered. In this paper, dissolved oxygen, COD, heavy metals, cyanide, formaldehyde and other elements were the main monitoring indicators.

The data transmission module is a bridge between the water quality monitoring module and the early warning module. Its task is to timely and accurately transmit the relevant data collected by the water quality monitoring module to the early warning module. At present, there are many different data transmission methods, each of which has its advantages and disadvantages. The commonly used remote transmission methods include wireless channel, dial-up telephone line, global mobile communication system, general packet radio service technology, etc. In practical application, the transmission mode of real-time monitoring data should be reasonably selected according to the actual situation. In this case, this paper used the general packet radio service technology to realize the data exchange between the water quality monitoring module and the water quality early warning module. After the simulation signal of the original data is collected by the water quality monitoring module, the simulation signal can be converted into a format that can be recognized by the computer through the signal converter, and the data can be transmitted to the early warning module through the data transmission system module.

The early warning module is the last part of water quality monitoring, which is responsible for storing and processing initial data. After receiving the information from the water quality monitoring module, it would be stored in the original storage system. Through statistics and analysis of the initial data, the water quality would be modeled using relevant methods to identify and forecast the changes in water quality, and the corresponding emergency response plan would be developed according to the actual situation [12-13].

3. Evaluation of Early Warning Module of Water Pollution Early Warning System

The early warning module of the water pollution early warning system includes data storage and data processing, as shown in Figure 2.

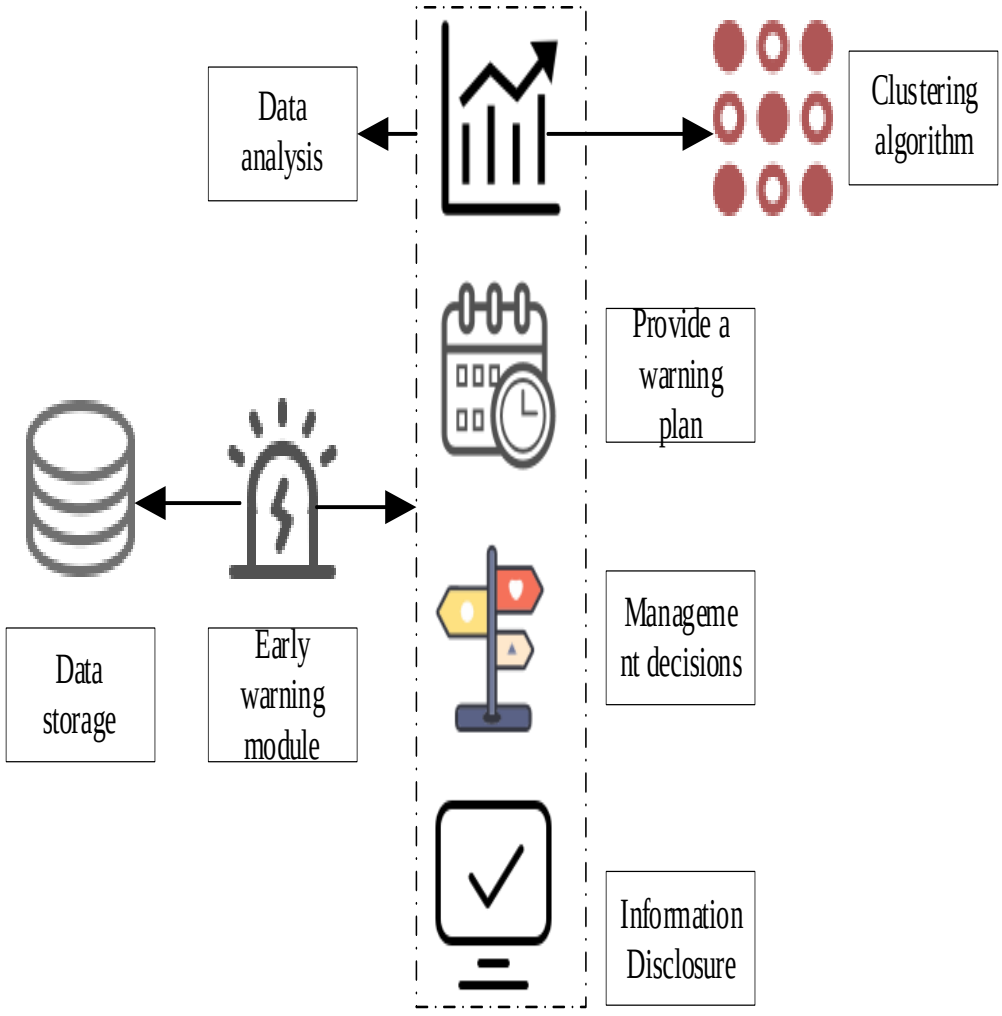


Figure 2. Early warning module of water pollution early warning system

The data storage part is responsible for receiving the information transmitted by the water quality monitoring module. It is based on the distributed data structure and can meet the requirements of data input and retrieval. The main function of the data processing part is to analyze and process the previous data in the data storage part, and obtain the corresponding water quality change rules. It establish the corresponding mathematical model, so as to predict the future water quality and timely handle the sudden water quality change [14].

The clustering algorithm in machine learning is used for data analysis to predict water quality

[15]:

The water quality monitoring data is standardized, and the water quality monitoring data is expressed in vector form:

$$a_p = \left(a_{p_1}, a_{p_2}, ..., a_{p_n}\right)^S \tag{1}$$

The average value of a sampled attribute is expressed as follows:

$$\overline{a}_q = \frac{1}{m}\sum_{i=1}^{m} a_{pq} \ , q \in [1, n] \tag{2}$$

The sampling standard deviation of a feature is calculated:

$$T_q = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(a_{pq} - \overline{a}_q\right)^2} \tag{3}$$

The water quality monitoring data shall be standardized by the following formulas:

$$a'_{pq} = \frac{a_{pq} - \overline{a}_q}{T_q} \tag{4}$$

$$a''_{pq} = \frac{a'_{pq} - \min\limits_{1 \le p \le m}\left(a'_{pq}\right)}{\max\limits_{1 \le p \le m}\left(a'_{pq}\right) - \min\limits_{1 \le p \le m}\left(a'_{pq}\right)} \tag{5}$$

If the similarity value is used to calculate the similarity of water quality monitoring samples, the following formulas can be obtained:

$$T_{pq} = \frac{1}{2}\left(d_{pq} + u_{pq}\right) \tag{6}$$

$$d_{pq} = 1 - \sqrt{\frac{1}{n}\sum_{c=1}^{n}\left(a_{pc} - a_{qc}\right)^2} \tag{7}$$

$$u_{pq} = \frac{\sum_{c=1}^{n} a_{pc} a_{qc}}{\sqrt{\sum_{c=1}^{n} a_{pc}^2 \sum_{c=1}^{n} a_{qc}^2}} \tag{8}$$

Among them, T_{pq} is the similarity between water quality monitoring sample p and water quality monitoring sample q.

The cluster file structure is used to search and obtain the cluster status of the cluster data. That is to say, the initial data is divided into several clusters. In this way, the cluster center of each cluster is an average vector of all sample vectors in the cluster, which can be obtained from the following formula:

$$\overline{a} = \frac{1}{n}\sum_{p=1}^{n} p \tag{9}$$

In this cluster, n is the number of water quality monitoring samples.

Each cluster is clustered and archived to form a cluster center file. Cluster search is mainly based on the cluster center file. After providing a cluster center file, each cluster is searched and compared with the information provided. On this basis, the similarity of each cluster is calculated, and the cluster with the greatest similarity is selected as the initial search. That is to say, the known data to be queried is included in this cluster. The software can be used to calculate the similarity between the query data vector and each cluster center, and obtain the maximum value after the calculation, so as to quickly determine which cluster the data to query is. By comparing the information to be checked with the information of its cluster, the information of which water quality monitoring site the information to be checked can be obtained. In addition, in order to ensure the accuracy of the retrieval results, the monitoring data should also be updated. At the same time, the data of different periods should be compared and analyzed, so as to better understand the evolution of the water quality of the water source, and better monitor and control the water quality.

Provision of warning plan: The warning plan is also the disposal plan for pollution accidents. This section contains many optimal treatment methods that should be adopted under various pollution conditions.

Management decision: When the environmental pollution event occurs or the predicted value reaches the warning line, the early warning module would analyze the characteristics of the environmental quality and compare it with the pollution situation of the monitoring plan. If there is a similar situation, the system can be used to deal with it directly, which greatly shortens the processing time and feeds back relevant information to relevant departments.

Information release: The water quality of the water source is regularly released, and the emergency is issued with early warning. When the format of the published information does not change much, the template can be used to input only text and automatically generate corresponding graphics.

In terms of ensuring the effectiveness of the water pollution early warning system, the quality of water quality monitoring should be strengthened from the following aspects. Before water quality monitoring, corresponding monitoring instruments and equipment shall be configured and maintained to ensure the normal operation of water quality monitoring and early warning system and avoid errors in monitoring data. In order to ensure the representativeness of samples, the collection, transportation and testing of water quality samples should strictly comply with relevant specifications and procedures. Sampling must be carried out in strict accordance with the specified method. The temperature of the water source sample should be maintained to ensure the stability of the water source sample. In the process of water source sample monitoring, the monitoring and early warning personnel are required to have a certain level of professional technology and can use the instrument correctly to ensure the standardization of the inspection work. The monitoring and early warning personnel shall strictly monitor the temperature and humidity of the laboratory and conduct standardized operation to avoid adverse effects on the monitoring and test results.

4. Effect and Recognition Evaluation of Water Source Pollution Early Warning System

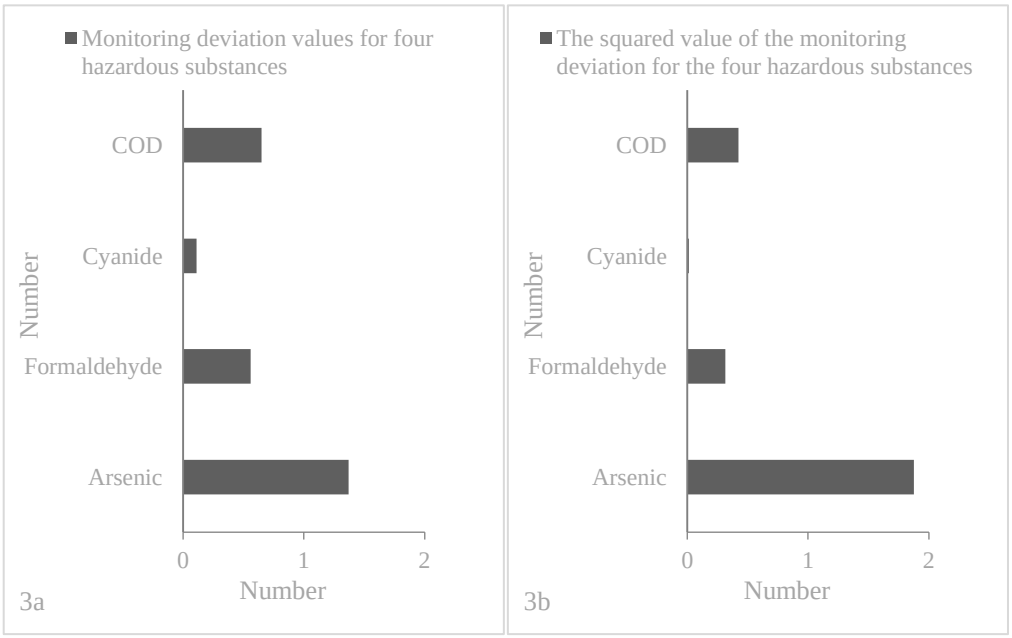
In this paper, the water pollution early warning system based on clustering algorithm was used to test arsenic, formaldehyde, cyanide and COD, so as to determine the monitoring and early warning accuracy of the water pollution early warning system. In the form of interviews, the recognition degree of the water pollution monitoring and early warning technology personnel to the water pollution early warning system proposed in this paper was investigated.

Arsenic, formaldehyde, cyanide and COD solutions of different concentrations were prepared and tested with the water pollution early warning system. The test results were shown in Table 1.

Table 1. The results of testing by the early warning system for water pollution

	True value of hazardous substances (mg/L)	The actual detection value of the hazardous substance (mg/L)
Arsenic	20	18.63
Formaldehyde	13.5	12.94
Cyanide	2	1.89
COD	100	99.35

The average value of absolute error and mean square error were used to evaluate the water source environmental pollution monitoring effect of the water source pollution early warning system. First, the monitoring deviation value and the square value of the monitoring deviation of the water source pollution early warning system for four hazardous substances were obtained, as shown in Figure 3.



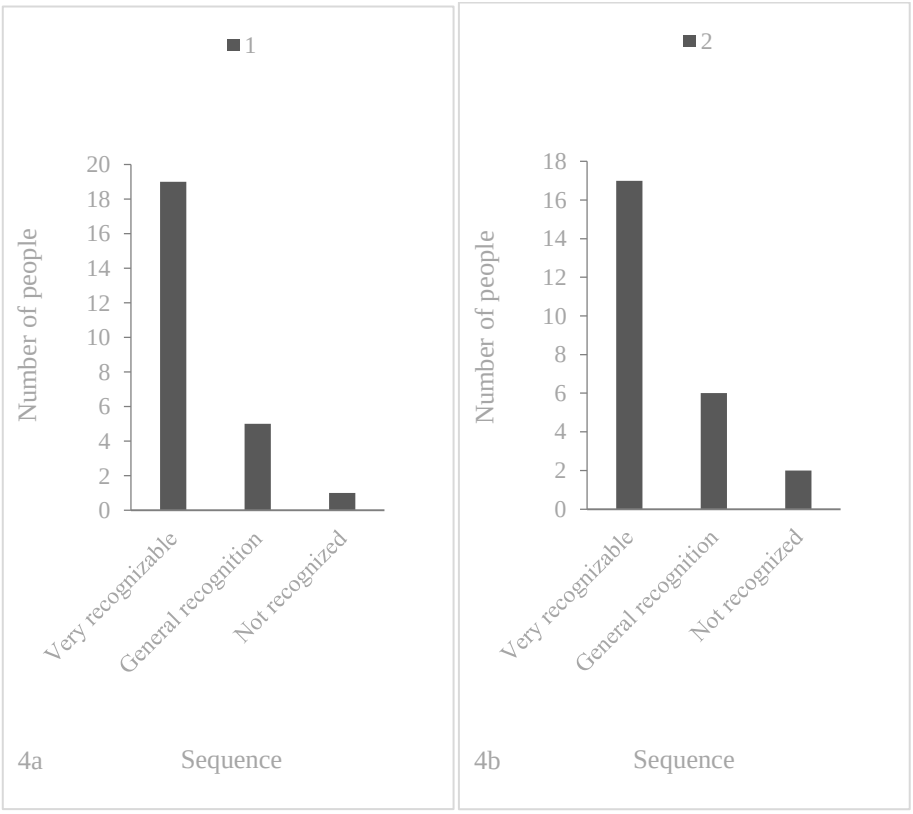
3a: Monitoring deviation value *3b: Squared value of monitoring deviation*

Figure 3. The monitoring deviation value and the square value of the monitoring deviation of the four harmful substances of the water pollution early warning system

As shown in Figure 3, Figure 3a showed the monitoring deviation of the water pollution early warning system for four hazardous substances, and Figure 3b showed the square value of the monitoring deviation of the water pollution early warning system for four hazardous substances. It could be seen from Figure 3a that the monitoring deviation of the water pollution early warning system for arsenic, formaldehyde, cyanide and COD was not high. Among them, the deviation value of arsenic monitoring by the water pollution early warning system was 1.37, and the deviation value of formaldehyde monitoring by the water pollution early warning system was 0.56. The deviation value of cyanide monitoring by the water pollution early warning system was 0.11, and the deviation value of COD monitoring by the water pollution early warning system was 0.65. It could be calculated from the above data that the average absolute error of the water source pollution early warning system in monitoring four hazardous substances was about 0.67. It could be seen from Figure 3b that the square value of the monitoring deviation of the source pollution early

warning system for arsenic, formaldehyde, cyanide and COD was also not high. Among them, the square deviation of arsenic monitoring by the water pollution early warning system was 1.8769, and the square deviation of formaldehyde monitoring by the water pollution early warning system was 0.3136. The square deviation of cyanide monitoring by the water pollution early warning system was 0.0121, and the square deviation of COD monitoring by the water pollution early warning system was 0.4225. It could be calculated from the above data that the mean square error of the water pollution early warning system in monitoring four hazardous substances was about 0.66.

This paper investigated the recognition of the water pollution monitoring and early warning system proposed in this paper by the water pollution monitoring and early warning technicians in the form of interviews. Two interviews were conducted. There were 25 people in each interview. The recognition of monitoring and early warning technicians to the water pollution early warning system was shown in Figure 4.



4a: Results of the first interview *4b: Results of the second interview*

Figure 4. Monitoring and early warning technicians' approval of the early warning system for water pollution

As shown in Figure 4, Figure 4a showed the recognition of monitoring and early warning technicians to the water pollution early warning system under the first interview scenario, and Figure 4b showed the recognition of monitoring and early warning technicians to the water pollution early warning system under the second interview scenario. From Figure 4a, the following information could be obtained: 19 monitoring and early warning technicians were highly recognized for the water pollution early warning system, and 5 monitoring and early warning technicians were generally recognized for the water pollution early warning system. One monitoring and early warning technician did not approve the water pollution early warning system. According to the interview, the monitoring and early warning technician believed that the network transmission

efficiency of the water pollution early warning system was not high enough. From Figure 4b, the following information could be obtained: 17 monitoring and early warning technicians were highly recognized for the water pollution early warning system, and 6 monitoring and early warning technicians were generally recognized for the water pollution early warning system. Two monitoring and early warning technicians did not approve the water pollution early warning system. According to the interview, among the monitoring and early warning technicians who did not recognize the water pollution early warning system, one thought that the pollutant detection types of the water pollution early warning system were not enough, and the other thought that the pollutant detection accuracy of the water pollution early warning system still needed to be improved.

5. Conclusion

Water is an essential factor for people's survival. Clean, sufficient and stable water supply is an important guarantee for the development of a city. With the acceleration of urbanization and industrialization, people are increasingly dependent on water. However, at the same time, there are water pollution problems in many areas. Based on this, this paper studied the monitoring and early warning of water quality pollution, and proposed a water quality monitoring and early warning system based on clustering algorithm. This paper introduced the three modules of water source sampling, data transmission and early warning included in the water source water quality monitoring and early warning system in detail, and focused on the early warning module, which also studied the effect of water source pollution monitoring and early warning. The following conclusions were drawn through the study: From the perspective of the mean absolute error and mean square error, the water source monitoring and early warning system based on clustering algorithm had high monitoring and early warning accuracy. Although the research in this paper could provide some reference for the related research of water source monitoring and early warning, its research also had limitations. Due to the limited access, it is hoped that it would have the opportunity to increase the sample of interviewees in future research.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Hossain, SM Zakir, and Noureddine Mansour. "Biosensors for on-line water quality monitoring–a review." Arab Journal of Basic and Applied Sciences 26.1 (2019): 502-518.

[2] Barzegar, Rahim, Mohammad Taghi Aalami, and Jan Adamowski. "Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model." Stochastic Environmental Research and Risk Assessment 34.2 (2020): 415-433.

[3] Echavarria-Caballero, Carolina. "Assessment of Landsat 5 images atmospherically corrected with LEDAPS in water quality time series." Canadian Journal of Remote Sensing 45.5 (2019): 691-706.

[4] Li, Z. H. O. U., Ling-zhi Li, and Ji-kun Huang. "The river chief system and agricultural non-point source water pollution control in China." Journal of Integrative Agriculture 20.5 (2021): 1382-1395.

[5] Khatri, Punit. "Towards the green analytics: Design and development of sustainable drinking water quality monitoring system for Shekhawati Region in Rajasthan." MAPAN 36.4 (2021): 843-857.

[6] Alam, Arif U. "Fully integrated, simple, and low-cost electrochemical sensor array for in situ water quality monitoring." ACS sensors 5.2 (2020): 412-422.

[7] Wu, Jianhua. "Statistical and multivariate statistical techniques to trace the sources and affecting factors of groundwater pollution in a rapidly growing city on the Chinese Loess Plateau." Human and Ecological Risk Assessment: An International Journal 26.6 (2020): 1603-1621.

[8] Chen, Zeng. "An adaptive data cleaning framework: a case study of the water quality monitoring system in China." Hydrological Sciences Journal 67.7 (2021): 1114-1129.

[9] Zhu, WeiYu. "Dynamic early warning method based on abnormal detection of water quality time series." Environmental Science & Technology (China) 41.12 (2018): 131-137.

[10] Pujar, Prasad M.. "Real-time water quality monitoring through Internet of Things and ANOVA-based analysis: a case study on river Krishna." Applied Water Science 10.1 (2020): 1-16.

[11] Xia, Xuemin. "Genetic algorithm hyper-parameter optimization using Taguchi design for groundwater pollution source identification." Water supply 19.1 (2019): 137-146.

[12] Lakshmikantha, Varsha. "IoT based smart water quality monitoring system." Global Transitions Proceedings 2.2 (2021): 181-186.

[13] Kamal, Noha. "Early Warning and Water Quality, Low-Cost IoT Based Monitoring System." JES. Journal of Engineering Sciences 47.6 (2019): 795-806.

[14] Thai-Nghe, Nguyen, Nguyen Thanh-Hai, and Nguyen Chi Ngon. "Deep learning approach for forecasting water quality in IoT systems." International Journal of Advanced Computer Science and Applications 11.8 (2020): 686-693.

[15] Haghiabi, Amir Hamzeh, Ali Heidar Nasrolahi, and Abbas Parsaie. "Water quality prediction using machine learning methods." Water Quality Research Journal 53.1 (2018): 3-13.