Scholar Publishing Group
Water Pollution Prevention and Control Project
https://doi.org/10.38007/WPPCP.2020.010106
ISSN 2790-2730 Vol. 1, Issue 1: 51-60

Time Series Data Cleaning and Early Warning of River Basin Water Quality

Kunst Rafael[*]

University Greifswald, Germany

[*]*corresponding author*

Keywords: Water Quality Time Series, Data Cleaning, Water Quality Early Warning, River Basin

Abstract: In order to effectively control the safety of water environment and promote the continuous improvement of water environment quality, all national environmental monitoring stations in the basin should conduct intensive monitoring of water quality, so that the environmental protection department can timely understand the water quality dynamics and make decisions on early warning and water quality analysis. How to monitor and warn watershed water quality data has become one of the current research hotspots. However, most water quality sensors often suffer from daily maintenance difficulties, database input errors and sensor measurement errors. Therefore, this paper analyzed the framework and process of data cleaning of time series, then optimized and analyzed the early warning model, and finally analyzed the early warning effect through abnormal early warning algorithm. The water quality warning effect after data cleaning was 15.6% higher than that before data cleaning, and the indicator detection effect was 11.2% higher than that before data cleaning. In short, data cleaning of time series and early warning model can improve water quality.

1. Introduction

Considering the increasingly serious situation of water resources security, effective water quality monitoring and early warning have aroused great interest in dealing with water pollution. Water quality monitoring is crucial for improving the protection and management of water resources at present. Relevant environmental services and facilities institutions have evaluated the lake in accordance with relevant laws and regulations, and evaluated it by monitoring the water quality of rivers and streams. Aquatic environment monitoring technology involves real-time monitoring of water quality indicators and determination of water quality changes and patterns in the analysis process to ensure the good state of the aquatic environment in the region.

Many scholars have studied and analyzed the water quality time series. Loc Ho Huu proposed a way forward to improve the performance of artificial intelligence (AI) models to better consider the periodicity of data. He explored a transfer function method, in which the water quality time series of a parameter was predicted based on the set of other parameters [1]. Raseman William J proposed a nonparametric time series method based on nearest neighbor self-weight sampling, which resampled historical data with “eigenvector” as the condition at a given time to generate values at a later time [2]. In order to verify the previously observed link between water safety programs and health outcomes, Setty Karen E used this time series study to examine the site-specific relationship between water related exposure and incidence rate of acute gastroenteritis in three locations in France and Spain [3]. Echavarria-Caballero Carolina applied the normalized differential water index to Landsat TM5 image to distinguish water and non-water information, and compared the surface reflectance with the field measured water reflectance [4]. Haghiabi Amir Hamzeh investigated the performance of artificial intelligence technology in predicting the water quality composition of the Tirei River in southwestern Iran [5]. Barzegar Rahim developed two independent deep learning models to predict the two water quality variables of Little Prespa Lake in Greece, namely dissolved oxygen and chlorophyll. The main novelty of this study was to establish a coupled neural network model to predict water quality variables [6]. Yang Deuk Seok, taking into account the operation of Changning-Haman Weir, which was located in the Nanjiang River flowing into the Luodong River, analyzed the pattern and trend of the water level and water quality of the Luodong River using the self-organization map and local weighted scatter map smoothing technology. The water quality of the Chixi Station of the Luodong River based on automatic monitoring was greatly affected by the Luodong River and the Nanhe River [7]. The above studies have described the specific role of water quality time series, but they are not involved in data cleaning.

Many scholars have analyzed and studied the prediction of water quality. Chen Zeng’s robust detection of patterns under low signal-to-noise ratio was a basic challenge to analyze high-frequency data, especially in water quality monitoring. He proposed an adaptive method based on empirical wavelet transform and multi-scale fuzzy entropy to achieve time series data cleaning [8]. Meng Qingxuan proposed a data cleaning method based on improved balanced iterative reduction and hierarchical clustering algorithm. He constructed the clustering feature tree of water quality data, and used the clustering method to obtain the clustering vector of the clustering feature tree [9]. Khatri Punit proposed the development of a sustainable water quality monitoring system, and considered the principle of green analysis when developing the proposed system to reduce the time consumption and labor cost of measurement [10]. Alam Arif U believed that the sensing system could be easily modified and programmed to integrate other sensors. This ability could be used to monitor a series of water quality parameters, demonstrating the integrated system for monitoring faucets, swimming pools and lake water. The system opened up possibilities for various low-cost and ubiquitous environmental monitoring applications [11]. Rao K Raghava put forward a skilled observation method for water level and water quality of overhead water tank to reduce current water waste and provide better water quality [12]. The above studies have described the early warning and prediction effects of water quality, but there are still some deficiencies in model optimization.

In order to study the effect of water quality time series data cleaning and early warning in river basins, this paper analyzed the technology of water quality data cleaning and the construction of early warning database, and then analyzed the characteristic curve according to the abnormal early warning algorithm. Through experimental comparison and analysis, it was found that the water quality early warning effect and indicator detection effect after water quality data cleaning have greatly improved compared with those before cleaning. Compared with other literatures, this paper focused on comparing the noise removal effect and missing filling effect before and after data cleaning.

2. River Basin Water Quality Time Series Data Cleaning Technology

2.1. Reasons for Dirty Data of River Basin Water Quality

In the technical application of water quality detection facilities, the water quality data collected by different types of sensors are often characterized by large amount of data, large coverage, long time, inconsistent detection indicators and other complex and changeable characteristics. Today, water quality monitoring data is generally input into the water quality database to meet the business requirements for site and management. Therefore, in the process of data collection and input, various problems may occur, such as data noise and data loss. The water resources early warning model helps to develop strategies to improve water safety in drought and other flood situations, and applies sensors to other areas to monitor water quality [13]. Observation sensors are also not used in areas with some geographical and financial constraints. Water quality data collection usually requires daily on-site maintenance and management, which involves many problems, such as sensor measurement error or low sensitivity. This may lead to insufficient noise level or water quality data, resulting in more dirty water quality data, as shown in Figure 1.

General noise data include high-frequency noise and Gaussian white noise, but traditional noise reduction methods cannot adaptively suppress different noise data. The lack of water quality data is mainly divided into three categories. One is the shortage of random data, which can be used to calculate the average value of adjacent points with the missing value. The second is short-term continuous missing. Machine learning model can be used to predict the additional value of missing filling. The last is large-scale continuous missing data.

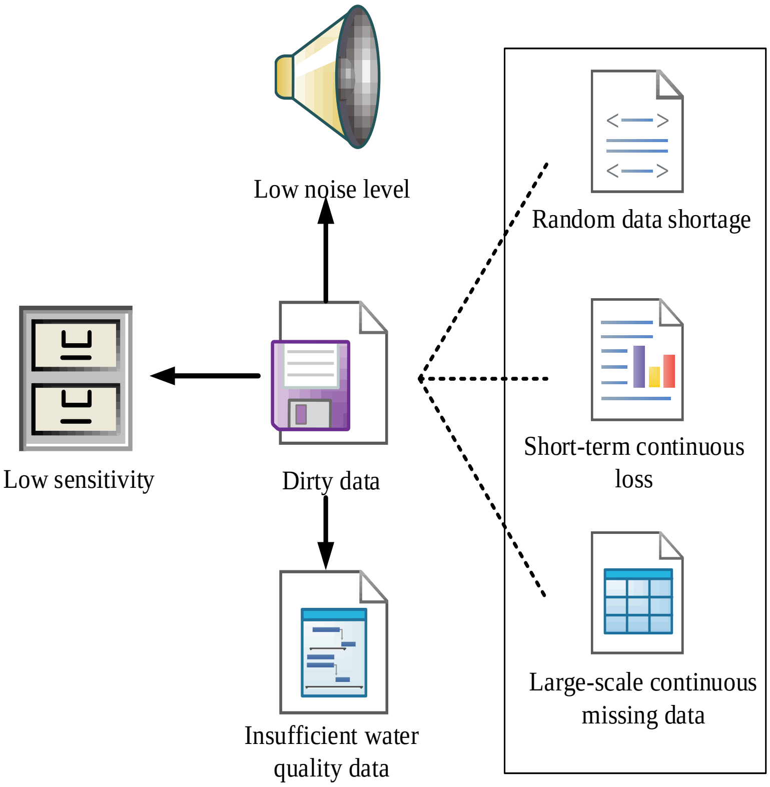

Low noise level
Random data shortage
Low sensitivity
Dirty data
Short-term continuous loss
Large-scale continuous missing data
Insufficient water quality data

Figure 1. Cause analysis of dirty water quality data

2.2. Framework and Process of Water Quality Time Series Data Cleaning

Data cleaning is usually the last process of detecting and correcting identifiable errors in data files, which helps to develop appropriate data cleaning strategies to effectively detect illegal data and generate high-quality time series data. First, the dirty data in the time series is analyzed, usually including five important steps, as shown in Figure 2. The first is noise anomaly data analysis, including thorough analysis of the dirty data contained in the time series (such as the original data set) to solve the problem of accidental and consistent data loss, and finally determine the appropriate purity control strategy. The second is to formulate data cleaning rules and strategies. Considering the analysis of time series of dirty data, this paper aims to formulate data cleaning strategies, algorithm rules and models applicable to data noise and data error. By detecting water quality anomalies, isolated forest algorithms are used, which can dynamically and accurately predict the time series of water quality parameters [14]. The third is to check the cleaning rules. The validation rules include verifying its validity and accuracy through simulation and test examples. If it is not verified, the cleaning policy, algorithm model parameters and other operations can be configured to maximize data cleaning performance. The fourth is to implement the cleaning process. The correct policy rules defined before cleaning are used to clean up dirty data. The fifth is to complete and high-quality data sets. Data cleaning (obvious characteristic trend, no default value) is waiting for further review and analysis.

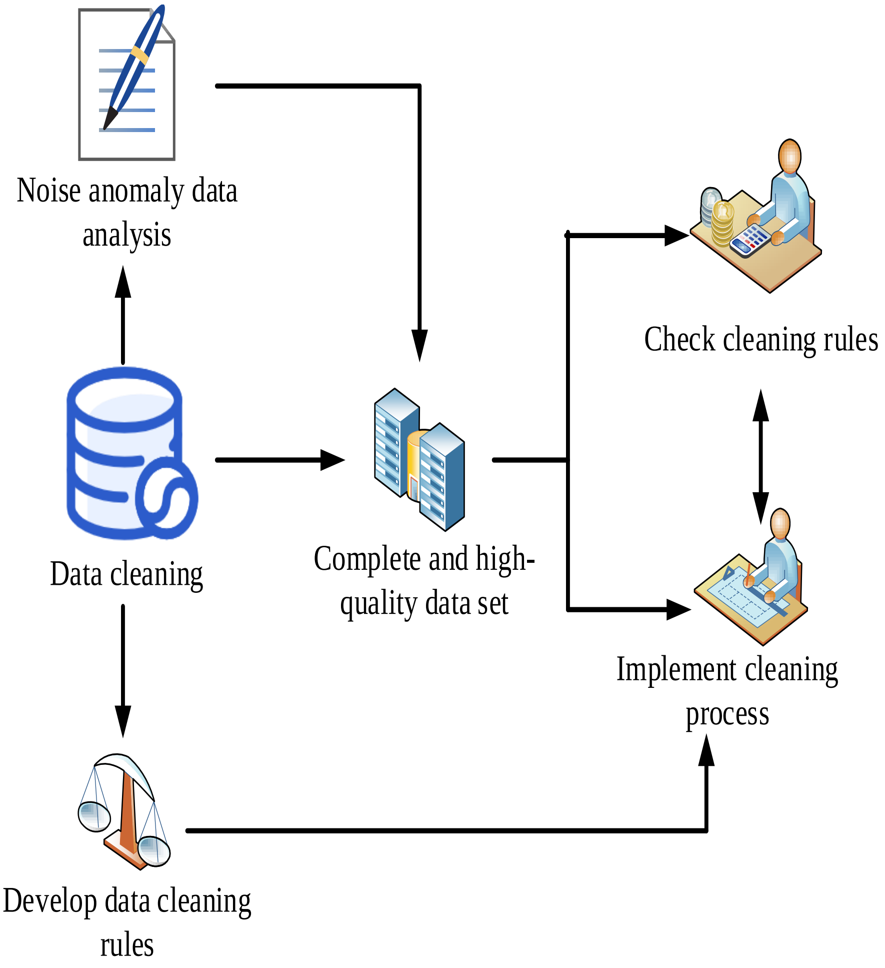

Noise anomaly data analysis
Check cleaning rules
Data cleaning
Complete and high-quality data set
Implement cleaning process
Develop data cleaning rules

Figure 2. Water quality time series data cleaning process

2.3. Key Technologies for Water Quality Time Series Data Cleaning

The key technology of water quality data cleaning is mainly divided into two stages: noise data detection and removal stage and missing data filling stage, as shown in Figure 3. In the stage of noise data collection and elimination, the water quality time series data are usually disturbed by high-frequency anomalies and Gaussian noise. Combined with water quality data, the algorithm combining empirical microwave transform and multi-scale fuzzy entropy threshold function is used to adjust to deal with all sounds that do not match the time distribution of water attributes. At the stage of filling in missing data, when the loss rate is low, traditional statistical methods or machine learning methods are used to predict missing data. However, at a higher loss rate, the existing algorithm used to fill in the lost data is invalid, because there is no strong correlation between the training samples around the lost data and the previous statistical data. Long-term and short-term memory models can better handle time series data and rely on time series data for a long time. Migration learning training is allowed to move between similar data fields.

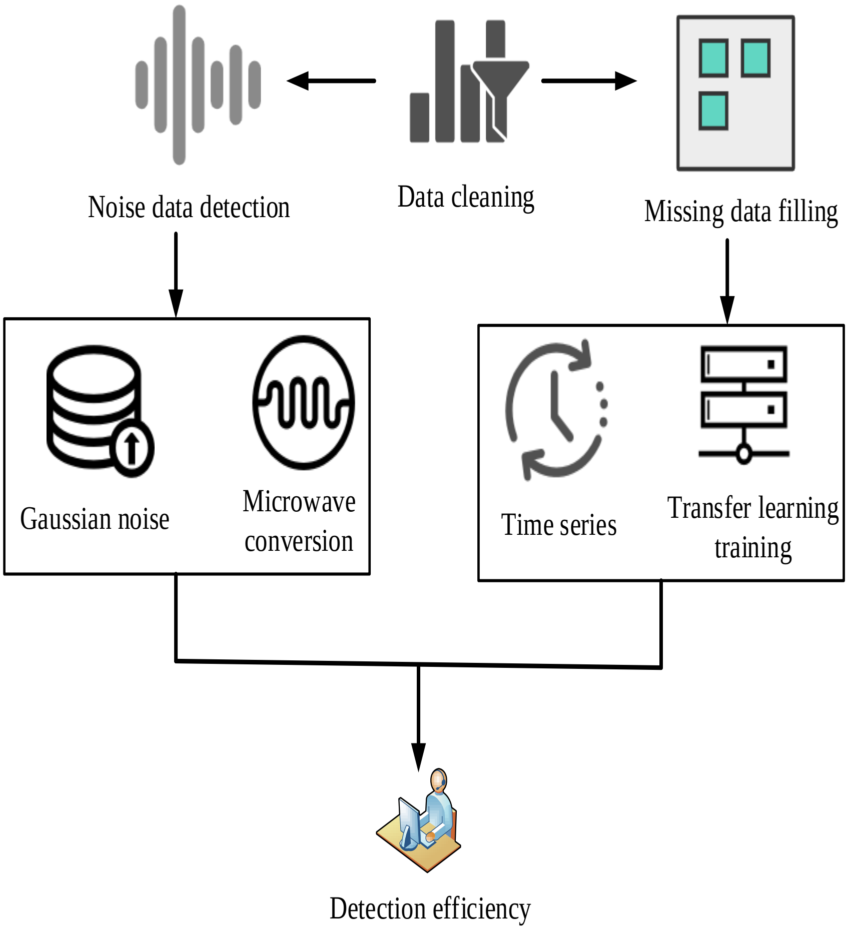


Figure 3. Key technologies for water quality data cleaning

3. Optimization of Early Warning Database of River Basin Water Quality Time Series

3.1. Database Expansion Design

In order to prevent database insertion, update, and problems related to deletion and other exceptions and duplication in the database, the design of relational database should meet the requirements of relationship standardization as far as possible. Because the database, especially the online monitoring data table, provides a large amount of data, the municipal database records millions of records every year, which depends on the actual amount of data stored in the database. If all online monitoring data are stored in the same table, the collected data has a significant impact on the performance of data requests. The data of recent months are usually used for practical applications such as query, estimation and prediction. The existing water quality monitoring plan includes comprehensive physical, chemical and biological analysis. It is still expensive and time-consuming to obtain these data on site [15]. According to the table separation principle, if each main process with high access frequency needs an unrelated subset of the table, the entire table can be separated. The speed of database query and search is considered to ensure the performance of data query. The online monitoring data table can be separated and saved by year. If years of data monitoring are required, the two tables can be viewed and connected. Restorage of redundant data can improve access speed, but the cost of maintaining data integrity increases. When updating related columns, triggers can be used to immediately update or store processes or applications.

3.2. Index Strategy

In the process of performance optimization, the selection of the columns to be indexed is one of the most important steps to optimize performance. If there is no index, the database must analyze the entire table in the first record to find the required record. In order to sort the commonly used columns in the query sentence, the database creates an index. Only one index can be used per query. Therefore, if each query can only be used for columns with search criteria, separate indexes can be set for these fields. If multiple fields are used as columns with search criteria at the same time, the index containing these fields must be created. Generally, it is necessary to find the site name after the site code or the site code after the site name, which can index the site code and site name columns in the water quality alarm database alarm system table. In the online monitoring data table, the data is usually the location code. It is filtered and queried according to the project code and monitoring time, and a complex index containing these three fields can be created. In the big data monitoring table, the effectiveness of index query is more obvious. The use of indexes can improve the efficiency of queries and also save costs. The more indexes in the database, the better it is. The indexes occupy disk space and need regular maintenance. For operations that need to write data, the index itself has been changed, resulting in slower data writing. Therefore, index creation is the balance between query speed and write speed, and corresponding indicators must be formulated based on the actual situation.

3.3. Database Backup of Water Quality Early Warning

The assessment, prediction, modeling and processing of water quality safety warning shall be carried out according to the water quality warning database. The database data damage or loss may be caused by human error, software error and hardware error. If a network error or computer virus occurs, it seriously affects the normal process of alarm analysis and decision-making. Therefore, the security and stability of the database is particularly important. The backup database must be established regularly. If a vulnerability occurs, the backup file can be used for recovery.

4. Evaluation of Early Warning System for Abnormal Water Quality in River Basins

In order to study the effect of water quality data cleaning and early warning in river basins, this paper analyzes water quality data through time series model, and constructs water quality data prediction equation. Then the true positive rate and false positive rate of the working characteristic curve of the water quality early warning model are studied. First, this paper calculates the water quality data prediction equation as follows:

$$\hat{y}(x,t)=(s_x+t)s_t \tag{1}$$

Among them, x is the water period and t is the interval of the water period. s_x is the seasonal factor of the water period, and s_t is the estimated mean value of the water period. Then, the water quality characteristic curve of the river basin is analyzed as follows:

$$c=\frac{p}{p+q} \tag{2}$$

$$d=\frac{j}{j+k} \tag{3}$$

c is the true positive rate. p is the actual exception and the detection exception. q is the actual abnormality, and the detection is normal. d refers to false positive rate, j refers to actual normal and abnormal detection. k is the actual normal and the detection is normal.

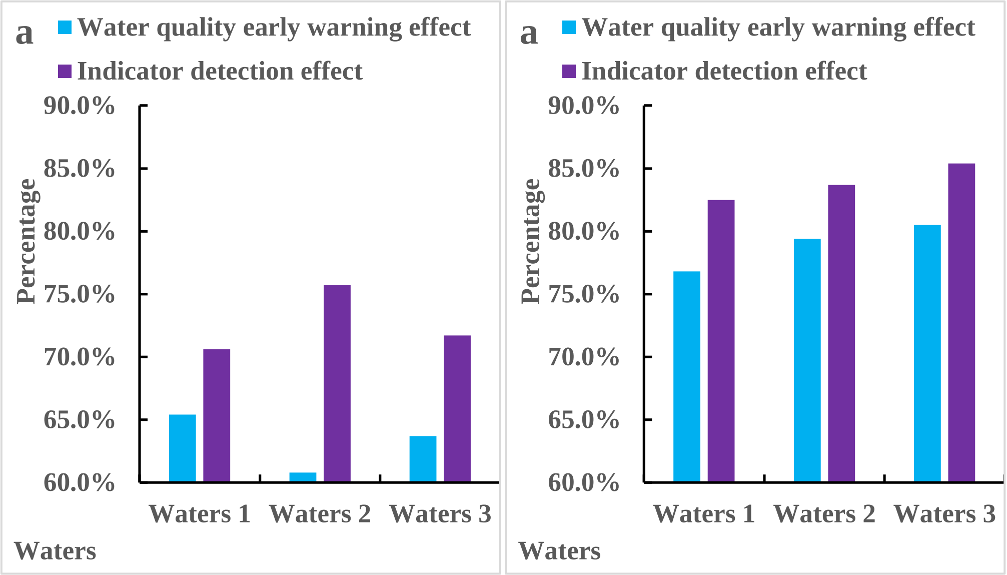
5. Water Quality Time Series Data Cleaning and Early Warning Model Experiment

In order to study the specific application effect of water quality time series data cleaning and early warning model in river basin, this paper investigated the noise removal effect and the missing filling effect of time series data cleaning, and then compared the water quality early warning effect and indicator detection effect before and after data cleaning. First of all, this paper investigated the noise removal effect and the missing filling effect of three water areas before and after data cleaning. The specific survey results are shown in Table 1.

Table 1. Noise removal effect and missing filling effect of three water areas before and after data cleaning

	Noise removal effect		Missing filling effect	
	Before data cleaning	After data cleaning	Before data cleaning	After data cleaning
Waters 1	0.52	0.71	0.34	0.82
Waters 2	0.48	0.75	0.30	0.88
Waters 3	0.50	0.79	0.37	0.81

According to the data described in Table 1, before data cleaning, the noise removal effect of water area 1 was 0.52, and the missing filling effect was 0.34. The noise removal effect of water area 2 was 0.48, and the missing filling effect was 0.30. The noise removal effect of water area 3 was 0.50, and the missing filling effect was 0.37. After data cleaning, the noise removal effect of water area 1 was 0.71, and the missing filling effect was 0.82. The noise removal effect of water area 2 was 0.75, and the missing filling effect was 0.88. The noise removal effect of water area 3 was 0.79, and the missing filling effect was 0.81. On the whole, the noise removal effect before data cleaning was 0.50, and the missing filling effect was 0.34. The noise removal effect after data cleaning was 0.75, and the missing filling effect was 0.84. Through comparison, it can be seen that the noise removal effect after data cleaning was 0.25 higher than that before data cleaning, and the missing filling effect was 0.50 higher than that before data cleaning. After data cleaning of the time series of water quality data, it can effectively improve the clarity of water quality, not only help fill in the missing detection data, but also reduce the noise of the data to reduce the interference caused by noise. Finally, the water quality early warning effect and indicator detection effect before and after data cleaning were analyzed, and a total of three waters were investigated. The specific investigation results are shown in Figure 4.



a. Before data cleaning *b. After data cleaning*

Figure 4. Water quality early warning effect and indicator detection effect before and after data cleaning

Figure 4a shows before data cleaning, and Figure 4b shows after data cleaning. It can be seen from Figure 4a that before data cleaning, the water quality warning effect of water area 1 was 65.4%, and the indicator detection effect was 70.6%. The water quality early warning effect of water area 2 was 60.8%, and the indicator detection effect was 75.7%. The water quality early warning effect of water area 3 was 63.7%, and the indicator detection effect was 71.7%. According to Figure 4b, after data cleaning, the water quality early warning effect of water area 1 was 76.8%, and the indicator detection effect was 82.5%. The water quality early warning effect of water area 2 was 79.4%, and the indicator detection effect was 83.7%. The water quality early warning effect of water area 3 was 80.5%, and the indicator detection effect was 85.4%. On the whole, the water quality early warning effect before data cleaning was 63.3%, and the indicator detection effect was 72.7%. After data cleaning, the water quality warning effect was 78.9%, and the indicator detection effect was 83.9%.

According to the experimental analysis, the water quality warning effect after data cleaning was 15.6% higher than that before data cleaning, and the indicator detection effect was 11.2% higher than that before data cleaning. The data cleaning of water quality time series not only improves the early warning effect of water quality, but also improves the detection accuracy of various indicators of water quality.

6. Conclusion

Most unusual water pollution events are unpredictable and irregular. According to the frequency of the event, multi-indicator water quality monitoring and water pollution anomaly early warning have a positive impact on maintaining the ecological balance of the water environment. Based on the water quality monitoring data and other relevant information, the water quality data processing and early warning technology platform can effectively early warning the sudden and gradual pollution of water quality, and effectively improve the technical level of water quality pollution control and the quality of regional water environment, thus ensuring the safety of drinking water, which realizes the early warning of abnormal pollutants in the water environment and provides feasible solutions.

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] Loc Ho Huu. Deep neural network analyses of water quality time series associated with water sensitive urban design (WSUD) features. Journal of Applied Water Engineering and Research. (2020) 8(4): 313-332. https://doi.org/10.1080/23249676.2020.1831976

[2] Raseman William J. Nearest neighbor time series bootstrap for generating influent water quality scenarios. Stochastic Environmental Research and Risk Assessment. (2020) 34(1): 23-31. https://doi.org/10.1007/s00477-019-01762-3

[3] Setty Karen E. Time series study of weather, water quality, and acute gastroenteritis at Water Safety Plan implementation sites in France and Spain. International journal of hygiene and environmental health. (2018) 221(4): 714-726. https://doi.org/10.1016/j.ijheh.2018.04.001

[4] Echavarria Caballero Carolina. Assessment of Landsat 5 images atmospherically corrected with LEDAPS in water quality time series. Canadian Journal of Remote Sensing. (2019) 45(5): 691-706. https://doi.org/10.1080/07038992.2019.1674136

[5] Haghiabi, Amir Hamzeh, Ali Heidar Nasrolahi, Abbas Parsaie. Water quality prediction using machine learning methods. Water Quality Research Journal. (2018) 53(1): 3-13. https://doi.org/10.2166/wqrj.2018.025

[6] Barzegar Rahim, Mohammad Taghi Aalami, Jan Adamowski. Short-term water quality variable prediction using a hybrid CNN-LSTM deep learning model. Stochastic Environmental Research and Risk Assessment. (2020) 34(2): 415-433. https://doi.org/10.1007/s00477-020-01776-2

[7] Yang Deuk Seok. Patterns and Trends of Water Level and Water Quality at the Namgang Junction in the Nakdong River Based on Hourly Measurement Time Series Data. Journal of Environmental Science International. (2018) 27(2): 63-74. https://doi.org/10.5322/JESI.2018.27.2.63

[8] Chen Zeng. An adaptive data cleaning framework: a case study of the water quality monitoring system in China. Hydrological Sciences Journal. (2020) 67(7): 1114-1129.

[9] Qingxuan Meng, Jianzhuo Yan. A data cleaning method for water quality based on improved hierarchical clustering algorithm. International Journal of Simulation and Process Modelling. (2019) 14(5): 442-451. https://doi.org/10.1504/IJSPM.2019.104120

[10] Khatri Punit. Towards the green analytics: Design and development of sustainable drinking water quality monitoring system for Shekhawati Region in Rajasthan. MAPAN. (2020) 36(4): 843-857.

[11] Alam Arif U. Fully integrated, simple, and low-cost electrochemical sensor array for in situ water quality monitoring. ACS sensors. (2020) 5(2): 412-422. https://doi.org/10.1021/acssensors.9b02095

[12] Rao K. Raghava. IOT based water level and quality monitoring system in overhead tanks. International Journal of Engineering & Technology. (2018) 7(2): 379-383. https://doi.org/10.14419/ijet.v7i2.7.10747

[13] Lishuo Guo, Lifang Wang. Construction and application of water security early‐warning model. Water and Environment Journal. (2020) 36(3): 458-468. https://doi.org/10.1111/wej.12778

[14] Weiyu Zhu. Dynamic early warning method based on abnormal detection of water quality time series. Environmental Science & Technology (China). (2018) 41(12): 131-137.

[15] Kamal Noha. Early Warning and Water Quality, Low-Cost IoT Based Monitoring System. JES. Journal of Engineering Sciences. (2019) 47(6): 795-806. https://doi.org/10.21608/jesaun.2019.115742