Scholar Publishing Group
Water Pollution Prevention and Control Project
https://doi.org/10.38007/WPPCP.2022.030405
ISSN 2790-2730 Vol. 3, Issue 4: 38-46

Water Pollution Early Warning Model Based on Decision Tree Algorithm and Electronic Information Technology

Jiaqing Li*

Philippine Christian University, Philippine

ljq@sxu.edu.cn

**corresponding author*

Keywords: Decision Tree Algorithm, Electronic Information Technology, Water Pollution, Early Warning Model

Abstract: After the water pollution (WP) incident, it has caused a serious impact on the natural environment, people's health, residents' daily life, and social economy. It is more urgent to promote the economy and environment with high level and quality, which puts forward new requirements for environmental managers. In recent years, the prediction and early warning of water bodies using decision tree (DT) algorithm has received extensive attention. Therefore, this paper adopts the DT algorithm and establishes a WP early warning model (EWM) with the support of electronic information technology (EIT). This model is used to monitor a river reach and water quality indicators, analyze the changes of water quality parameters, and analyze the water quality status. This efficient and accurate prediction and early warning technology is conducive to timely response to WP events, which is of great significance for ecological protection and environmental decision-making.

1. Introduction

In recent years, China's government has attached importance to water pollution control and committed to raising people's awareness of environmental protection, so the quality of the water environment has been greatly improved, but the rapid development of industry has led to frequent water pollution accidents, how to prevent water pollution or minimise the damage caused by water pollution accidents has become one of the hot spots of research in the field of water environment.

Scholars have achieved certain research results in water pollution early warning. For example, some scholars have constructed a new model for water quality prediction based on BP neural network optimisation and used the method to identify the time and location of pollution sources in rivers and lakes, verifying the excellent performance of the method. In addition, combining the

inverse probability method and linear regression model, a differential evolutionary algorithm was used to identify the location of pollution sources, the time of discharge and the total discharge of unexpected water pollution events, transforming the pollution source identification problem into a more solvable optimisation model [1]. Some scholars have implemented groundwater inherent vulnerability assessment based on GIS technology and groundwater pollution early warning using set-pair analysis [2]. Some researchers have developed a GIS-based groundwater early warning system in C++, providing a water pollution early warning method for GIS, evaluating and grading pollution, considering the current situation and development trend of water pollution early warning, and carrying out the development of water pollution early warning and water quality dynamic mapping output function [3-4]. At this stage for water quality data processing methods are often still on the traditional calculation software, hope that in the future can be more advanced technology to monitor water quality.

This paper first introduces the related concepts and models of DT algorithm, and expounds the characteristics of WP accidents. Then, based on the data number algorithm and EIT, a WP EWM is constructed, as well as the four main functional modules of the analysis model. Finally, the application of the EWM in water quality monitoring and water quality index prediction is analyzed.

2. Related Algorithms and WP Accidents

2.1. DT Algorithm

DTs are widely used and are commonly used to solve two-class classification problems. With regard to classification, we study known training data, generate a mapping template combining classification results and features, and use this template to map the classified data [5]. The information gain in a DT can be expressed in terms of information entropy. Information entropy is used to measure the similarity of samples. The lower the information entropy, the more similar the data in the branch and the higher the purity of the data.

$$Ent(D) = -\sum_{k=1}^{y} p_k \log_2 p_k \tag{1}$$

Where D denotes the sample set and the proportion of the kth class of samples is p_k. If the feature a can be taken to any value in, then the sample can be divided into V parts according to the feature a. The samples in each branch take the same value on the a feature, and this branch can be noted as DV, and Ent(DV) can be calculated according to equation (1), and then the information gain of the feature a is deduced based on the increase of the weight of the proportion as shown in equation (2) is shown.

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{\left|D^V\right|}{D} Ent(D^V) \tag{2}$$

$$Ent(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{3}$$

The DT classification method starts from the root node, starts testing for each feature, assigns each feature to a child node according to the test obtained, and calculates its information gain separately for the classification, so as to select the most efficient child node and recursively until it

reaches the leaf node [6].

2.2. Characteristics of WP Accidents

WP accidents are divided into sudden accidents and gradual accidents, both of which have their own distinctly different characteristics [7]. Sudden WP accidents in the accident occurred in the time and space, watershed sex, the impact of the consequences, disposal and impact time and emergency subjects and other general water pollution accidents have distinctly different characteristics.

Uncertainty in the waters in which they occur: waters are divided into reservoirs, lakes, rivers, estuaries, oceans and groundwater, but the velocity and flow of water have a significant impact on the rate of dispersion of pollutants [8]; even within the same river, different points of flow have very different flow characteristics.

Uncertainty about pollution sources: It is difficult to determine the pattern of pollution impacts and the capacity for environmental damage because of the different types and quantities of pollutants that cause water pollution incidents. However, the type and quantity of pollution sources are important parameters for dealing with water pollution incidents and mathematical modelling of water quality [9-10].

Uncertainty of harm: the exploitation of water in various forms and degrees, i.e. water functions differently [11]. Water pollution incidents of the same scale and degree of pollution also have different pollution consequences; if urban and rural water pollution incidents are of the same scale and degree, the consequences of the damage can be very different, with urban damage being much greater than rural damage and the negative impacts of urban water pollution being higher than rural [12-13].

3. DT Algorithm and EIT Based on the WP EWM

3.1. DT Algorithm and EIT in the Application of WP Early Warning

With the further development of artificial intelligence, the application of EIT to the information processing of water quality spectra can effectively improve the measurement stability, real-time and accuracy of traditional instruments. EIT is the primary technology for the construction of many intelligent systems in today's society, and the application of this technology in WP monitoring is to analyze water quality parameters through spectral instruments [14].

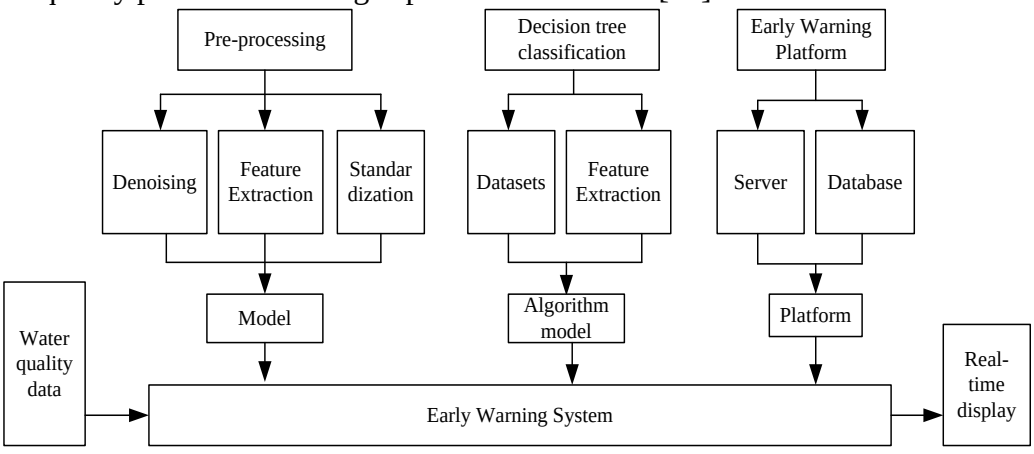

Figure 1. WP early warning method system structure diagram

As in Figure 1, after collecting water quality, the UV appearance was standardised using water quality data from the spectra to understand the water quality collection, thus enabling the transfer of early warning models, reducing the resources required to calibrate equipment and improving the use of models, resulting in a standardised UV-Vis spectral water quality database for spectral standard instruments [15]. Due to the standardised UV levels of different rivers, various indicators and algorithmic models were selected during data collection, so that a DT algorithm can be used to extract UV-Vis spectral water quality data features, in order to learn and model the spectral information used to extract spectral water quality data features, apply the classification of spectral information to study the quality of water spectral information, and analyse water quality through a decision algorithm classification [16]; using the water quality analysis model, based on the advantages of electronic information technology, and take full advantage of its efficient computing power and its portability, to achieve the processing of water quality data, and by setting up a water pollution early warning system environment, can real-time display of water quality, so as to achieve water pollution early warning [17-18].

3.2. The Main Modules of the WP Early Warning Platform

(1) Real-time data module

The system waits as simply as possible for information on the quality of the water module given in real time, these monitoring stations are evenly distributed in the basin and basically reflect the overall water quality in the basin. The interface shows the latest trends in the concentration of monitoring parameters, such as monitoring stations and real-time concentration values for different water quality parameters. A small box in the left corner of the system page provides information through alarms. When the parameters are based on standards or sudden water pollution, an option box showing alarm information and alarm sounds will be opened.

(2) Historical data module

The historical data module of the system has three sub-modules: single station history, multi-site history and data statistics. Only one station module can display station information at different times. After selecting a page, click the desired button to view the start time and the water quality parameters from the last time the selected page was set up to view the data in the default distribution format. If you want to see the difference in water quality parameters in the basin over a certain period of time, you can click on the graph button on the page.

(3) Water quality warning module

The system's water quality early warning module consists of two major components: water pollution early warning and threshold early warning. In the threshold warning module, the first water quality warning parameters should be set for the first time, when the water quality parameters measured with monitoring instruments are drawn in the centre of the set water quality parameter concentration curve, the alarm is set to issue a warning sound and the real water quality parameter changes are recorded. The process of water pollution warning is to first input the relevant hydrological and geographical information of the watershed where the water pollution occurs, and then selects the appropriate warning model, after setting the initial parameters of the model to study the algorithm and get the final result of the warning.

(4) Auxiliary function module

The auxiliary function module of the early warning system has valuable, easy to access and use assistive technologies. The module mainly consists of registration and login functions, and user rights setting functions. The login function secures the system by not allowing users to log in to the platform without a registered account, to ensure that registered users have safer access to system functions. Role and permission settings are the user can access the system functions to do a more

detailed boundary, ordinary users can only see real-time monitoring interface, past data interface and water quality assessment interface and other non-functional interface, while the administrator can enter information such as water quality assessment model, water quality warning mode data, through the permission control to ensure that the system can be long-term stable operation.

4. WP EWM Application

In order to verify the establishment of the WP EWM, selected a river section of the real-time online monitoring station data for application. Monitoring the site of dissolved oxygen, ammonia nitrogen, potassium permanganate, total phosphorus (TP) and other water quality indicators, they will be used as the verification object, the time range to choose nearly a year of observed data, monitoring frequency of 4 hours once, the basic characteristics of the data sequence is shown in Table 1 and Figure 2.

Table 1. Measured data series characteristic value (mg / L)

	Dissolved oxygen	Ammonia nitrogen	Potassium permanganate	TP
Max	14.68	17.41	23.25	1.36
Min	0.76	0.35	5.77	0.18
Mean	5.27	6.32	10.49	0.43
Variance	2.94	3.61	3.23	0.22
Median	5.38	4.73	11.05	0.40

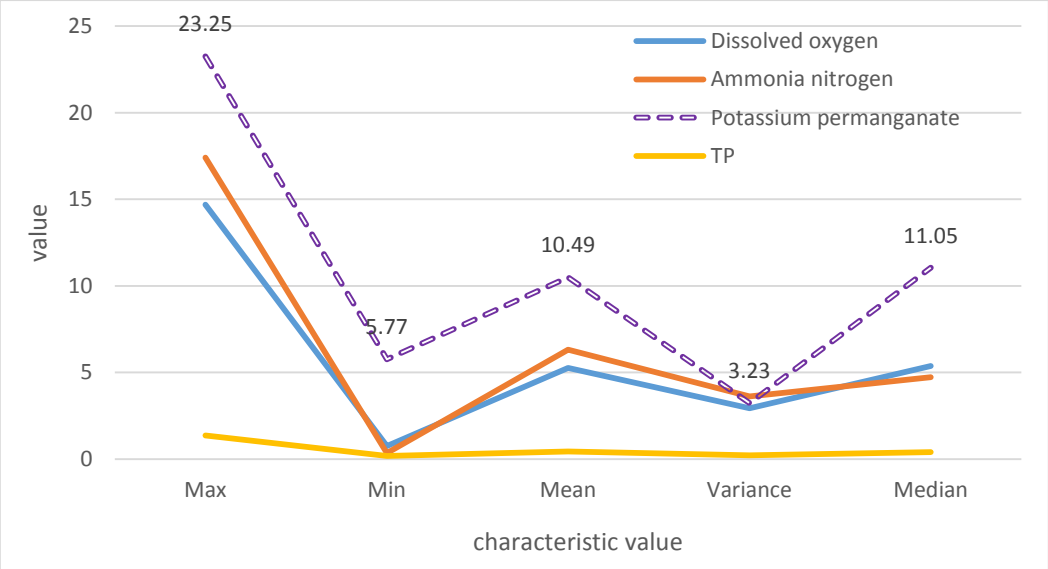

Figure 2. Distribution of evaluation indicator characteristics

As shown in Table 2 is the predicted and measured values of the four water quality indicators, the comparison shows that the difference between the predicted and actual values of each parameter is very small, indicating that the EWM in this paper has a very accurate prediction effect and can monitor the health of water quality.

Table 2. Comparison of predicted and measured values of water quality indicators

Monitoring serial number		30	40	50	60	70	80
Dissolved oxygen	Predicted value	3.58	6.14	2.73	4.92	3.36	5.95
	Measured value	2.46	6.23	2.75	4.87	3.42	5.91
Ammonia nitrogen	Predicted value	1.15	1.38	1.54	1.13	2.65	2.84
	Measured value	1.23	1.36	1.55	1.17	2.66	2.80
Potassium permanganate	Predicted value	5.93	7.48	8.12	3.74	6.49	3.61
	Measured value	5.96	7.45	8.09	3.68	6.53	3.63
TP	Predicted value	0.35	0.47	0.32	0.51	0.28	0.37
	Measured value	0.34	0.43	0.33	0.54	0.26	0.39

The prediction error results of the four water quality parameters are shown in Fig. 3. From the overall prediction effect of the four parameters, the WP EWM established in this paper shows relatively good prediction results, among which the prediction effect of TP and ammonia nitrogen is the best, the RMSE of TP is 0.38%, MSE is only 0.003%, MAE is 0.29%; the RMSE of ammonia nitrogen is 2.94%, MSE is only 0.08% and MAE was 2.26%.

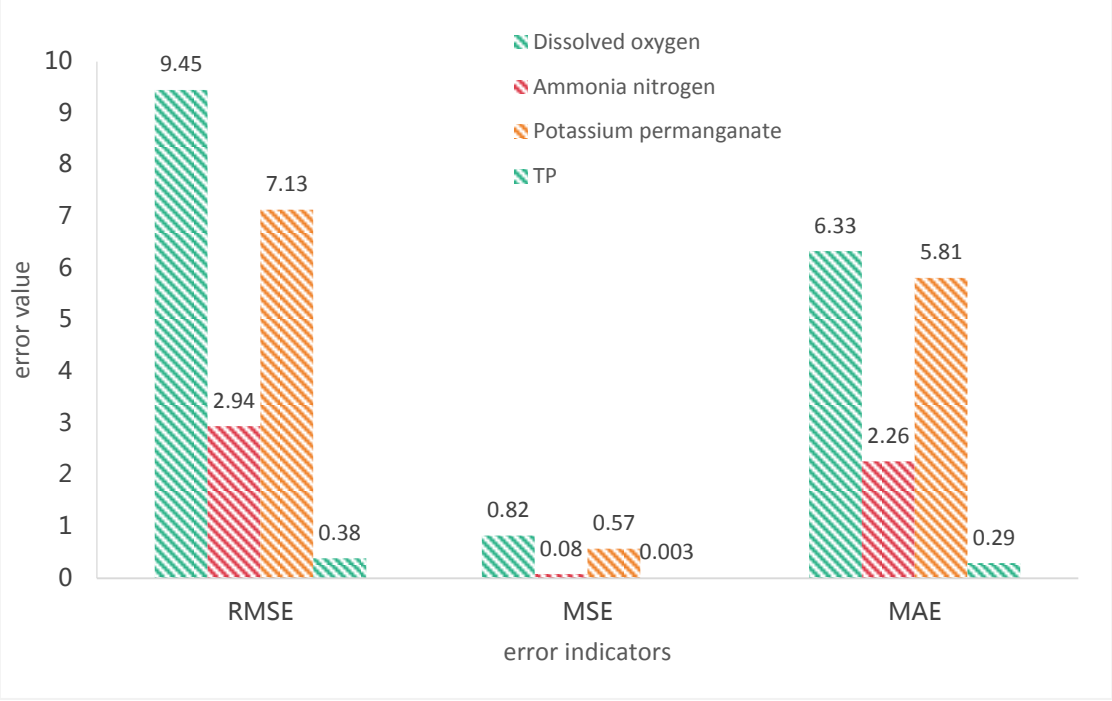

Figure 3. Four water quality parameters prediction results evaluation (%)

For the four water quality parameters of the predicted residual series, according to the minimum actual threshold method, respectively, calculated to find the corresponding minimum actual threshold, see Figure 4, the minimum actual threshold determined by the residual series, the

monitoring value within the threshold range in 99%. The abnormal thresholds of the four water quality parameters are 0.03, 0.26, 0.61, 0.043.

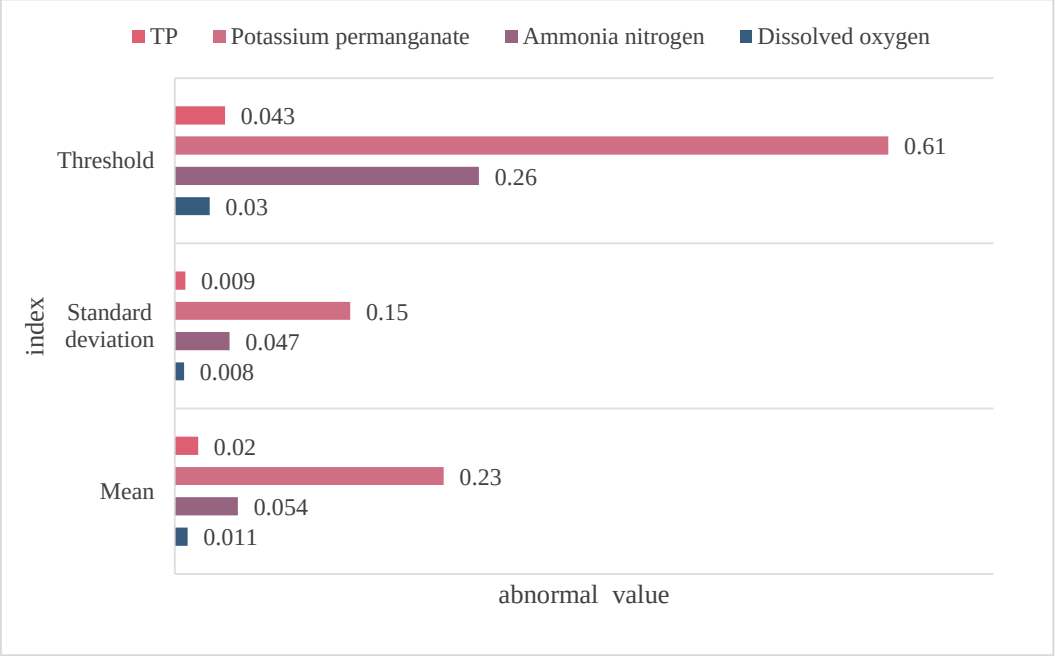

Figure 4. Abnormal thresholds for water quality parameters

In summary, the application of WP EWMs to monitor the water quality of water bodies and predict changes in water quality indicators can respond to WP events.

5. Conclusion

WP early warning system can predict the occurrence of sudden WP events and reduce the risk of water bodies being polluted, efficient and accurate early warning is essential for environmental decision-making. Current WP early warning is often by determining whether the indicators exceed a fixed threshold or range of change in the range of early warning, some abnormal events are difficult to be identified at an early stage, which increases the time from the occurrence of the accident to the treatment, but also increases the risk of environmental pollution. In this study, based on the real-time prediction of water quality, relying on EIT, timely and effective judgment of the abnormal phenomenon of the water body, and to determine the level of early warning, reduce the harm caused by sudden WP events, to provide a theoretical basis for WP disposal decisions.

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] Pascale Carayon, Megan E. Salwei. Moving toward a sociotechnical systems approach to continuous health information technology design: the path forward for improving electronic health record usability and reducing clinician burnout. J. Am. Medical Informatics Assoc. (2021) 28(5): 1026-1028. https://doi.org/10.1093/jamia/ocab002

[2] Anthony F. Breitzman. The relationship between web usage and citation statistics for electronics and information technology articles. Scientometrics. (2021) 126(3): 2085-2105. https://doi.org/10.1007/s11192-020-03851-5

[3] Marziye Narangifard, Hooman Tahayori, Hamid Reza Ghaedsharaf, Mehrdad Tirandazian: Early diagnosis of coronary artery disease by SVM, DT algorithms and ensemble methods. Int. J. Medical Eng. Informatics. (2022) 14(4): 295-305. https://doi.org/10.1504/IJMEI.2022.123921

[4] Chandrashekhar Azad, Bharat Bhushan, Rohit Sharma, Achyut Shankar, Krishna Kant Singh, Aditya Khamparia. Prediction model using SMOTE, genetic algorithm and DT (PMSGD) for classification of diabetes mellitus. Multim. Syst. (2022) 28(4): 1289-1307. https://doi.org/10.1007/s00530-021-00817-2

[5] Ferdinand Bollwein, Stephan Westphal. A branch & bound algorithm to determine optimal bivariate splits for oblique DT induction. Appl. Intell. (2021) 51(10): 7552-7572. https://doi.org/10.1007/s10489-021-02281-x

[6] Firoozeh Karimi, Selima Sultana, Ali Shirzadi Babakan, Shan Suthaharan. Urban expansion modeling using an enhanced DT algorithm. Geolnformatica. (2021) 25(4): 715-731. https://doi.org/10.1007/s10707-019-00377-8

[7] Vinay Arora, Rohan Singh Leekha, Inderveer Chana. An Efficacy of Spectral Features with Boosted DT Algorithm for Automatic Heart Sound Classification. J. Medical lmaging Health Informatics. (2021) 11(2): 513-528. https://doi.org/10.1166/jmihi.2021.3287

[8] Swati Chopade, Hari Prabhat Gupta, Rahul Mishra, Preti Kumari, Tanima Dutta. An Energy-Efficient River Water Pollution Monitoring System in Internet of Things. IEEE Trans. Green Commun. Netw. (2021) 5(2): 693- 702. https://doi.org/10.1109/TGCN.2021.3062470

[9] Amal Agarwal, Lingzhou Xue. Model-Based Clustering of Nonparametric Weighted Networks With Application to Water Pollution Analysis. Technometrics. (2020) 62(2): 161-172. https://doi.org/10.1080/00401706.2019.1623076

[10] Alda Henriques, Milton Fontes, Ana S. Camanho, Giovanna D'Inverno, Pedro Amorim, Jaime Gabriel Silva. Performance evaluation of problematic samples: a robust nonparametric approach for wastewater treatment plants. Ann. Oper. Res. (2022) 315(1): 193-220. https://doi.org/10.1007/s10479-022-04629-z

[11] Mohamed S. Abdalzaher, M. Sami Soliman, Sherif M. EI-Hady, Abderrahim Benslimane, Mohamed Elwekeil: A Deep Learning Model for Earthquake Parameters Observation in IoT System-Based Earthquake Early Warning. IEEE Internet Things J. (2022) 9(11): 8412-8424. https://doi.org/10.1109/JIOT.2021.3114420

[12] Natalia Jorquera-Bravo, Andrea Teresa Espinoza Perez, Oscar C. Vasquez. Toward a sustainable system of wastewater treatment plants in Chile: a multi-objective optimization approach. . Ann. Oper. Res. (2022) 311(2): 731-747. https://doi.org/10.1007/s10479-020-03777-4

[13] Ana S. Camanho, Flavia Barbosa, Alda Henriques. A system-level optimization framework for efficiency and effectiveness improvement of wastewater treatment plants. Int. Trans. Oper. Res. (2022) 29(6): 3370-3399. https://doi.org/10.1111/itor.13129

[14] K. Pavendan, V. Nagarajan. Modelling of wastewater treatment, microalgae groWPh and harvesting by flocculation inside photo bioreactor using machine learning technique. J. Intell. Fuzzy Syst. (2022) 43(5): 5607-5620. https://doi.org/10.3233/JIFS-212676

[15] Alima Chaouche, Ali Zemouche, Messaoud Ramdani, Khadidja Chaib Draa, Cedric Delattre. Unknown input estimation algorithms for a class of LPV/nonlinear systems with application to wastewater treatment process. J. Syst. Control. Eng. (2022) 236(7): 1372-1385. https://doi.org/10.1177/09596518221083729

[16] Neda Gorjian Jolfaei, Bo Jin, Leon van der Linden, Indra Gunawan, Nima Gorjian. A reliability-cost optimisation model for maintenance scheduling of wastewater treatment's power generation engines. Qual. Reliab. Eng. Int. (2022) 38(1): 2-17. https://doi.org/10.1002/qre.2956

[17] Alam Nawaz, Amarpreet Singh Arora, Wahid Ali, Nikita Saxena, Mohd Shariq Khan, Choa Mun Yun, Moonyong Lee. Intelligent Human-Machine Interface: An Agile Operation and Decision Support for an ANAMMOX SBR System at a Pilot-Scale Wastewater Treatment Plant. IEEE Trans. Ind. Informatics. (2022) 18(9): 6224-6232. https://doi.org/10.1109/TII.2022.3153468

[18] Imen Baklouti, Majdi Mansouri, Ahmed Ben Hamida, Hazem Numan Nounou, Mohamed N. Nounou. Enhanced operation of wastewater treatment plant using state estimation-based fault detection strategies. Int. J. Control. (2021) 94(2): 300-311. https://doi.org/10.1080/00207179.2019.1590735