Scholar Publishing Group
Water Pollution Prevention and Control Project
https://doi.org/10.38007/WPPCP.2022.030303
ISSN 2790-2730 Vol. 3, Issue 3: 19-27

The Characteristics Clustering of Lake Water Pollution Based on Machine Learning Algorithm

Holger Böse*

Chandigarh University, India

**corresponding author*

Keywords: Machine Learning Algorithm, K-Means Clustering, Lake Water Quality, Pollution Characteristics

Abstract: Water is the source of life and the material basis for human survival, so the healthy state of water resources and environment is of great significance to the orderly development of human society. However, at present, the pollution of lake water body is getting worse, causing serious damage to the lake water environment. In order to improve the lake water quality, this paper uses machine learning algorithm to sample and detect the water quality indicators of Lake Q, uses K-means clustering to analyze the maximum, minimum and average values of water quality, and compares and analyzes eight indicators, including TN, TP, NH3-N, CODMn, DO, BOD, WT and pH value, in the high water season The change trend of water quality concentration and the spatial distribution characteristics of water quality in the normal and dry seasons found that the lake water quality(LWP) was seriously polluted in the normal season, followed by the high water period(HWP) and the low water period(LWP). Through the analysis of the characteristics of lake water pollution, this paper learned the situation of lake water pollution, and provided prevention and control suggestions for solving the problem of lake water pollution.

1. Introduction

The adequacy of water resources may determine the state of development of a place's economic development. Having abundant water resources can lead to the progress of a region's civilisation, but in recent years, as population growth and people's needs for spiritual and material life increase, water resources are often exploited and used in various forms, leaving the water environment facing an unpromising trend of change. Although the exploitation of water resources has brought some benefits to economic development, people have also paid a high economic price in order to restore

Copyright: © 2022 by the authors. This is an Open Access article distributed under the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (https://creativecommons.org/licenses/by/4.0/).

water ecosystems and deal with water pollution.

Research on water pollution characterisation of lake water quality has yielded good results. For example, in a study of phosphate release mechanisms in sediments, some researchers found that, unlike mineral-bearing substances in microorganisms, organic phosphorus in water can degrade at a certain sunlight temperature and then release large amounts of dissolved phosphate, which can be used as nutrients for aquatic plants during their growth period [1]. A comprehensive organic index evaluation method was used to assess the level of nitrogen and phosphorus pollution in the water column in a study of a well-known lake lagoon, and the results showed moderate nutrient pollution in the surface sediments [2]. Some scholars used AHP grey correlation analysis to study four main sections of a river, and concluded that the main pollution indicators in the section were total phosphorus and ammonia nitrogen, and that water quality was better in the northern part of the entire stream section than in the southern part [3]. The study of water quality pollution characteristics of lakes can improve the water quality of lakes, which is a guide to improve the regional environmental quality and promote the harmonious development of ecological environment to a certain extent.

This paper first introduces the concept of machine learning, proposes a machine learning algorithm - K-means clustering algorithm, expounds the algorithm model, then proposes the role of K-means clustering in water quality analysis, then analyzes the water pollution characteristics of Q Lake in dry season, normal season, and wet season, and finally proposes water pollution prevention measures for the protection of water resources of Q Lake.

2. Introduction to Relevant Algorithms

2.1. Machine Learning Algorithm

Machine learning is an interdisciplinary subject covering probability theory, statistics, intelligent computing and other theories of knowledge that are important components of artificial intelligence. In recent years, resources for intelligent computing and cloud computing have expanded rapidly, the rich resources of huge networks of data can be fully shared, and machine learning methods have been widely used [4]. The aim of machine learning is to use large sample data to discover data containing attributes through appropriate algorithms, in creating data models with analytical capabilities, and then to quickly, efficiently and accurately classify or predict new sample data [5-6]. When the classification or prediction results for new sample data differ significantly from the actual control results, the parameters in the data model can be adjusted to optimise the model prediction performance so that the model produces results closer to the actual values. Figure 1 illustrates a basic process of machine learning.

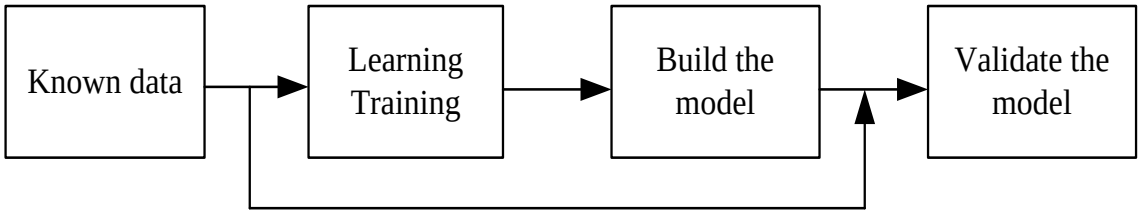

Figure 1. Machine learning process

The two basic elements of machine learning are learning algorithms and learning models, usually associated with data collection training to obtain classification or regression models through learning algorithms. In general, a complete machine learning process usually consists of the

following steps.

Engineering problem analysis problem analysis: reasoning and simplification of complex problems such as classification or regression in machine learning; data acquisition and analysis: obtaining better quality data types through parameterisation and informatisation of the model, especially in the case of classification problems, where category data should be shared as much as possible through data filtering or refinement; data pre-processing: this process includes data supplementation, removal of outliers, noise reduction processing, normalisation [7-8].

2.2. Clustering Algorithm

Clustering analysis is primarily used to identify and cluster groups of data. The use of data groups usually assumes that the same set of data features are similar and that there is a greater nature of differentiation between data features for different groups. This approach allows in principle to determine precisely which observations are similar so that groups of data with different characteristics can be delineated [9]. In the absence of variation in clustered responses, it is instead an unnecessary method, meaning that it attempts to find links between n variations in the absence of variation in response training. Since the possible classifications are not precisely defined, there are multiple methods of cluster analysis [10]. There are two most common methods of clustering: hierarchical clustering and partitioned clustering.

For hierarchical grouping methods, each clustering annotation is based on a combination of groupings based on a specific grouping algorithm until all clusters are grouped. For partitioned pooling, the number of k clusters is first determined, then observations are randomly assigned to k clusters and finally clustered into that cluster. k-means algorithms are most commonly used to cluster partitions [11].

K-means clustering is the most common classification algorithm for partitioning a given data set into k groups, and it is by far the most commonly used classical algorithm for clustering tasks and is suitable for the analysis of larger sample groupings [12]. The central concept of clustering algorithms is to classify the original dataset and evaluate the distance as a similarity metric through hierarchical iterations. The clustering results allow each of the generated categories to be compressed into the same class and allow the different classes to be independent of each other [13].

The K-means cluster center is represented by the distance between data points xi and xj. For any data point xi, define the local density ρ_i:

$$\rho_i = \sum_{j \in I_S \backslash \{i\}} \beta(d_{ij} - d_n) \tag{1}$$

$$\beta(x) = \begin{cases} 1, x < 0 \\ 0, x \geq 0 \end{cases} \tag{2}$$

Where d_n denotes the truncation distance, which is given by the user in advance, d_{ij} denotes the distance between samples i and j, and β is a parameter.

$$d(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{im} - x_{jn})^2} \tag{3}$$

$d(x_i, x_j)$ is the Euclidean distance between individuals of the K-means algorithm variables.

3. Water Pollution Feature Clustering of Lake Q Based on K-Means Algorithm

3.1. Application of K-Means Clustering in Water Quality Analysis

K-means cluster analysis is an important tool for trend prediction for data classification within many disciplines of research and is now widely used in environmental engineering, geographic mapping and administrative analytics. This paper uses K-means cluster analysis to analyse water quality monitoring indicators and the location of monitoring sections, using SPSS software for clustering calculations. K-means cluster analysis is a classification method that describes and defines data identifiers, and then carefully classifies and combines similar data. The description options of the most discrete variables, such as average value, standard deviation, maximum value and minimum value, are used for system clustering, and each monitoring indicator is selected as the analysis variable of LWP, FWP and HWP.

3.2. Analysis of Water Pollution Characteristics

Table 1. Water quality index parameter values

	DO	TP	TN	NH_3-N	COD_{Mn}	BOD	WT	pH
Max	13.42	8.2	10.5	15.64	7.97	9.86	23.2	7.9
Min	3.65	1.3	1.8	2.03	0.52	0.71	11.7	6.8
Mean	7.94	4.5	6.3	6.42	3.16	3.32	15.6	7.2
STD	1.83	0.08	0.74	1.17	0.25	0.43	3.27	0.15

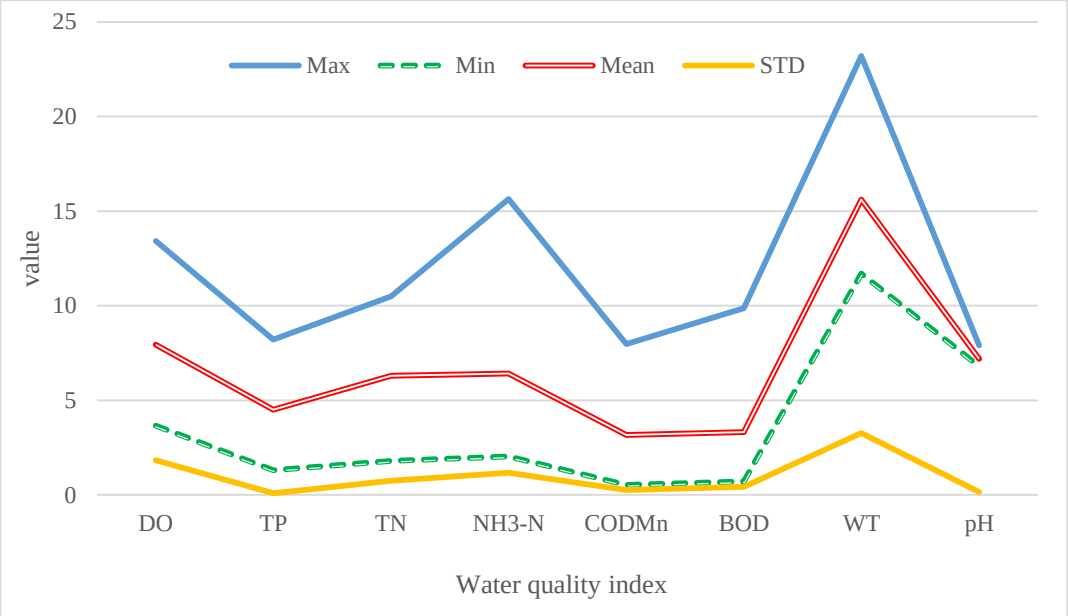

Figure 2. Changes in water quality in Lake Q

Statistical analysis of water quality indicators of Lake Q is shown in Table 1 and Figure 2. The maximum values of DO and NH_3-N are higher than other indicators, and the average and minimum values of COD_{Mn} are the smallest among these indicators.

3.3. Time Characteristics of Dry, Wet and Normal Periods

Table 2. Change of pH and Water Temperature

	LWP	FWP	HWP
pH	7.6	7.5	7.1
WT	13.4	19.2	20.8

The water temperature (WT) and the dry, normal and wet data of pH value are shown in Table 2. WT is the highest in the wet season, and its average value is 20.8 ℃ due to the seasonal temperature change. The water temperature in the wet season is often closely related to the eutrophication degree of the water body in the basin and the propagation of algae. When the temperature rises to 25-30 ℃, the water temperature will rise with it, which will lead to the massive propagation of cyanobacteria. The WT in the normal water period is relatively low, with an average value of 19.2 ℃. The dry season is also a cold time of the year, so the water temperature is only 13.4 ℃. The data of pH in three periods are 7.6 in dry season, 7.5 in normal season and 7.1 in wet season, showing weak alkalinity. The values of the three are within the standard range. At present, all the factories around Lake Q have to have special sewage treatment links. Few high pH or low pH sewage is directly discharged into the river body, so the change of pH value is mainly affected by different water levels in three periods, that is, by the amount of different river water volume in three periods.

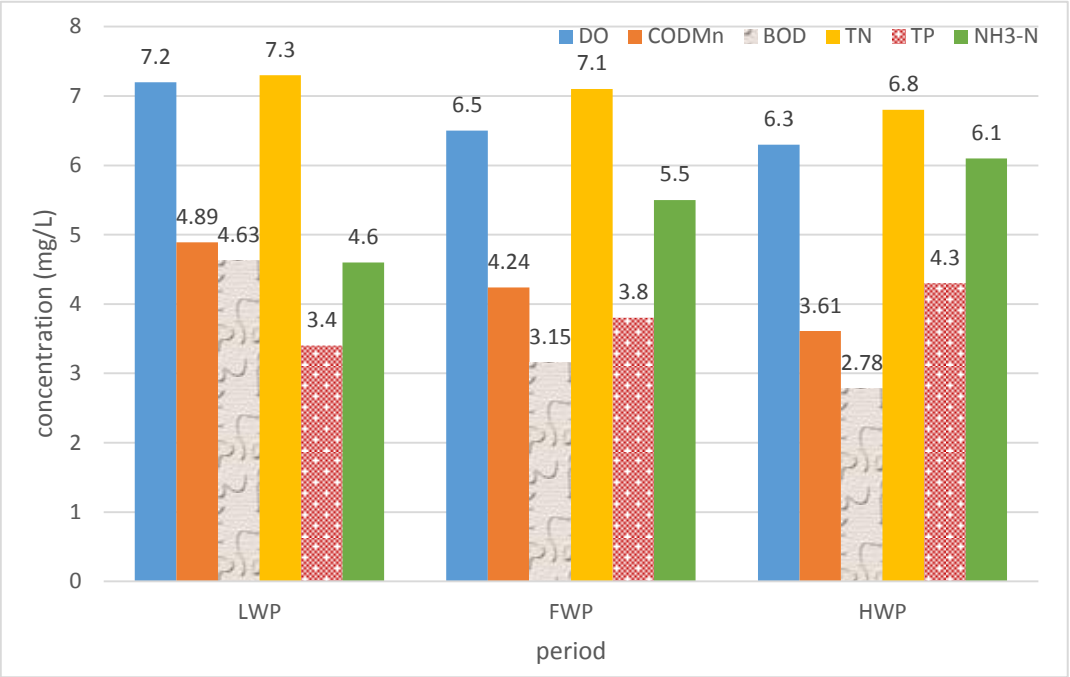

Figure 3. Change characteristics of DO, CODMn, BOD, TN, TP, NH3-N

The changes of nutrient and organic matter indicators DO, COD_{Mn} and BOD in Lake Q in the three water periods are shown in Figure 3. The content of DO is the highest in dry season and winter, because the temperature in Lake Q is the lowest and the pressure is the highest in this period. In addition, the algae hardly grow in this period, so the oxygen in the water is not consumed. The DO index reaches the Class I water standard in dry season, and the Class II water standard in wet season and calm season. In the wet season, on the contrary, the algal proliferation and low air pressure during this period lead to a large amount of oxygen in the water. The changes of COD_{Mn} and BOD

in three periods are not very large, and the values are relatively close. COD_{Mn} reaches Class II water standard in high water season. BOD reaches Class III water standard in dry season.

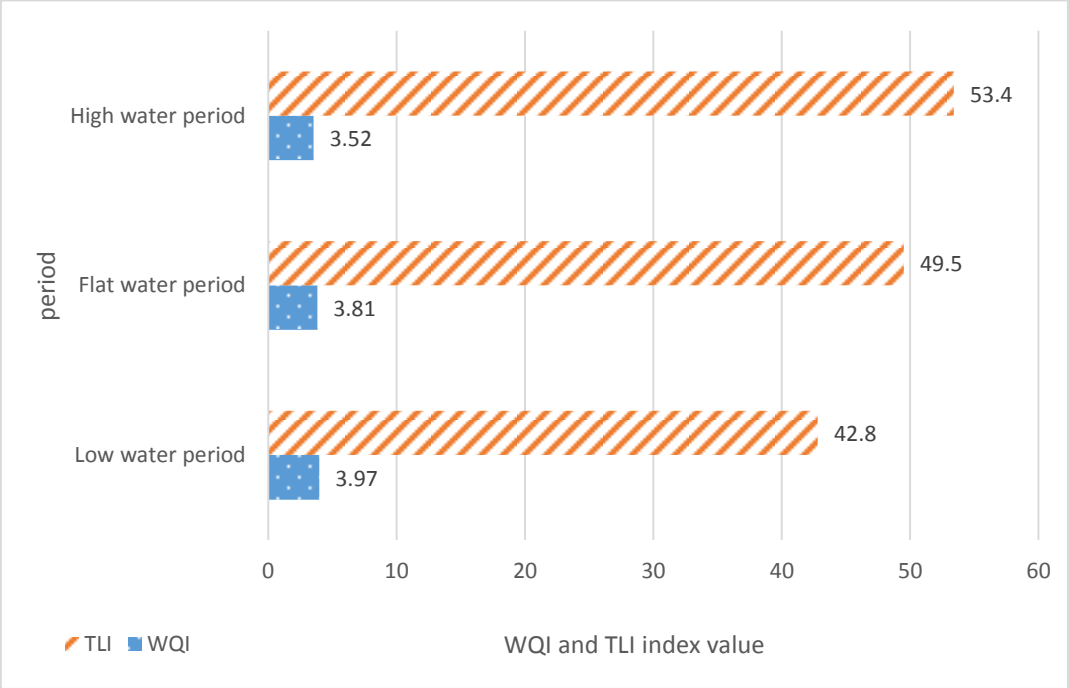

Figure 4. WQI and TLI changes in three periods

See Figure 4 for the changes of comprehensive water quality index (WQI) and Integrated nutrient index of Lake Q in the three periods. The WQI index in the three periods exceeded 3, all of which were in the range of Class III water indicators. The TLI was not in the state of eutrophication in the dry season and even in the wet season, while it exceeded the limit of 50 mild eutrophication in the wet season. The mean value of TLI was in a state of low eutrophication.

4. Prevention and Control of Lake Water Pollution

4.1. Strictly Implement "Zero Emission"

Due to the specific aquatic environment, the capacity of the lake's water resources is relatively small and projects with high water consumption and use do not meet environmental protection requirements in their production processes. Therefore, a strict 'zero discharge' strategy is required for water-intensive production projects. The development of new energy and materials industries, the setting of discharge standards for production wastewater, the creation of high recovery rates and the strengthening of wastewater control are important ways to achieve 'zero discharge' [14].

4.2. Strengthen Pollution Control of Existing Enterprises

As far as existing enterprises are concerned, we should strengthen supervision and management of enterprise effluent discharge standards, establish wastewater treatment facilities, prohibit illegal discharges and leakage discharges, and prohibit enterprises that do not decisively comply with industrial water use policies. We should actively promote advanced water treatment technologies, strengthen monitoring and inspection of wastewater treatment plants, and use public participation mechanisms to monitor and take measures to control corporate discharges where appropriate [15].

4.3. Vigorously Develop Clean Production and Encourage Reuse of Reclaimed Water

Enterprises should be encouraged to implement cleaner production techniques, reduce the amount of water used in the production process, improve water recycling rates and, where necessary, take water management measures to encourage the implementation of effective water conversion. Water reuse technologies and facilities should be vigorously promoted, and enterprises should be required to treat their own industrial wastewater. Effluent discharge permits and comprehensive control policies must be fully implemented to prevent excess effluent discharge. Regularly urge enterprises to carry out cleaner production, improve the level of cleaner production and promote cleaner production technologies to enterprises [16-17].

4.4. Continuously Improve the Quality and Efficiency of Environmental Management

The Ministry of Environmental Protection should continue to strengthen institutions, create employment opportunities, strengthen reasonable equipment, and set up special environmental supervision, inspection and monitoring departments to regularly inspect lake pollution sources and implement comprehensive lake water pollution supervision [18]. Regularly monitor water quality and establish a monitoring management system geared towards the lake water environment to deter discharge enterprises and eliminate environmental violations through excessive pollution discharge regulations. Effectively implement an environmental impact assessment system from the very beginning of enterprise production to avoid environmental violations at the root and source.

5. Conclusion

Lake water quality is an important indicator for studies to determine changes in lake water quality and to assess the health of lakes. Industrial wastewater, agricultural wastewater and human domestic sewage discharged into rivers and lakes increase the level of pollutants in the water environment of rivers and lakes by altering the hydrological characteristics of the water bodies, and many natural water systems are replaced by artificial water systems. The pollution of lake water quality not only threatens the normal survival of aquatic animals and plants, but also changes the basic functions of the lake ecosystem, directly affecting the daily production and life of residents living in the surrounding cities and counties, and also posing potential hazards to people's health. This paper uses clustering to analyse the characteristics of water pollution in lakes, so as to stimulate people's awareness of protecting the quality of the lake water environment, avoiding further deterioration of lake water quality and restoring the lake ecosystem is of great significance.

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] Angelika Zube, Dominik Kleiser, Alexander Albrecht, Philipp Woock, Thomas Emter, Boitumelo Ruf, lgor Tchouchenkov, Aleksej Buller, Boris Wagner, Ganzorig Baatar, Janko Petereit. Autonomously mapping shallow water environments under and above the water surface. Autom. (2022) 70(5): 482-495. https://doi.org/10.1515/auto-2021-0145

[2] Amandeep S. Gill, Stefan Germann. Conceptual and normative approaches to Al governance for a global digital ecosystem supportive of the UN Sustainable Development Goals (SDGs). Al Ethics. (2022) 2(2): 293-301. https://doi.org/10.1007/s43681-021-00058-z

[3] Ersan Batur, Derya Maktav. Assessment of Surface Water Quality by Using Satellite lmages Fusion Based on PCA Method in the Lake Gala, Turkey. IEEE Trans. Geosci. Remote. Sens. (2019) 57(5): 2983-2989. https://doi.org/10.1109/TGRS.2018.2879024

[4] Kai Matsui, Yoichi Kageyama, Hiroshi Yokoyama. Analysis of Water Quality Conditions of Lake Hachiroko Using Fuzzy C-Means. J. Adv. Comput. Intell. Intell. Informatics. (2019) 23(3): 456-464. https://doi.org/10.20965/jaciii.2019.p0456

[5] Samuel A. Ajila, Chung-Horng Lung, Anurag Das. Analysis of error-based machine learning algorithms in network anomaly detection and categorization. Ann. des Telecommunications. (2022) 77(5-6): 359-370. https://doi.org/10.1007/s12243-021-00836-0

[6] Koushiki Dasgupta Chaudhuri, Bugra Alkan. A hybrid extreme learning machine model with harris hawks optimisation algorithm: an optimised model for product demand forecasting applications. Appl. Intell. (2022) 52(10): 11489-11505. https://doi.org/10.1007/s10489-022-03251-7

[7] Felix Duong, Michael Gadermayr, Dorit Merhof, Christiane Kuhl, Philipp Bruners, Sven H. Loosen, Christoph Roderburg, Daniel Truhn, Maximilian F. Schulze-Hagen. Automated major psoas muscle volumetry in computed tomography using machine learning algorithms. Int. J. Comput. Assist. Radiol. Surg. (2022) 17(2): 355-361. https://doi.org/10.1007/s11548-021-02539-2

[8] Lamiaa M. Elshenawy, Chouaib Chakour, Tarek A. Mahmoud. Fault detection and diagnosis strategy based on k-nearest neighbors and fuzzy C-means clustering algorithm for industrial processes. J. Frankl. Inst. (2022) 359(13): 7115-7139. https://doi.org/10.1016/j.jfranklin.2022.06.022

[9] V. Parimala, K. Devarajan. Modified fuzzy C-means and K-means clustering based spectrum sensing using cooperative spectrum for cognitive radio networks applications. J. Intell. Fuzzy Syst. (2022) 43(3): 3727-3740. https://doi.org/10.3233/JIFS-212863

[10] Peter O. Olukanmi, Fulufhelo V. Nelwamondo, Tshilidzi Marwala, Bhekisipho Twala: Automatic detection of outliers and the number of clusters in k-means clustering via Chebyshev-type inequalities. Neural Comput. Appl.34(8): 5939-5958 (2022). https://doi.org/10.1007/s00521-021-06689-x

[11] Sharanpreet Kaur, Satwinder Singh: COVID-19 Infection and Air Pollution Characteristics in USA. Int. J. Appl. Evol. Comput.12(2): 16-35 (2021). https://doi.org/10.4018/IJAEC.2021040102

[12] Totan Garai, Harish Garg : Possibilistic multiattribute decision making for water resource management problem under single-valued bipolar neutrosophic environment. Int. J. Intell. Syst. 37(8): 5031-5058 (2022). https://doi.org/10.1002/int.22750

[13] Suresh Muthulingam, Suvrat S. Dhanorkar, Charles J. Corbett : Does Water Scarcity Affect Environmental Performance? Evidence from Manufacturing Facilities in Texas. Manag. Sci. 68(4): 2785-2805 (2022). https://doi.org/10.1287/mnsc.2021.4013

[14] Tristan Thielmann : Environmental conditioning: Mobile geomedia and their lines of becoming in the air, on land, and on water. New Media Soc. 24(11): 2438-2467 (2022). https://doi.org/10.1177/14614448221122190

[15] V. A. Miklush, I. A. Sikarev, Tatiana M. Tatarnikova: Organization of Environmental Monitoring of the Port Water Area by Processing an Anti-Interference Signal from a Vessel Traffic Control System. Autom. Control. Comput. Sci. 55(8): 999-1004 (2021) https://doi.org/10.3103/S0146411621080204

[16] Robert Smail, Christopher Donaldson, Rafael Govaerts, Paul Rayson, Carly Stevens : Uncovering Environmental Change in the English Lake District: Using Computational Techniques to Trace the Presence and Documentation of Historical Flora. Digit. Scholarsh. Humanit.36(3): 736-756 (2021). https://doi.org/10.1093/llc/fqaa047

[17] Mario Arzamendia, Daniel Gutierrez-Reina, Sergio Toral, Derlis Gregor, Eleana Asimakopoulou, Nik Bessis : Intelligent Online Learning Strategy for an Autonomous Surface Vehicle in Lake Environments Using Evolutionary Computation. IEEE Intell. Transp. Syst. Mag.11(4): 110-125 (2019). https://doi.org/10.1109/MITS.2019.2939109

[18] 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.5(2): 693-702 (2021). https://doi.org/10.1109/TGCN.2021.3062470