

COMPUTER SYSTEM FOR FORECASTING WATER QUALITY PARAMETERS BASED ON MACHINE LEARNING

Yue Zheng¹, Jianjun Wang², Sergiy Ryzhkov³,
Oksana Nechai⁴, Andrii Topalov^{5✉}, Oleksii Zivenko⁵,
Serhii Babchuk⁶, Taras Harasymiv⁶

Received on August 15, 2024

Presented by Ch. Roumenin, Member of BAS, on October 29, 2024

Abstract

The article presents research on environmental forecasting using machine learning methods for water quality analysis in the Southern Bug river. It focuses on the application of binary logistic regression and random forest regression to estimate the effects of various environmental parameters on key water quality indicators. The study involved data preparation by reading and processing CSV files with economic and environmental indicators and generating feature matrices and target vectors for model training. Both models were trained using nine years of data. The results can be used to develop water resource management strategies, predict the ecological state of water basins, and assess the influence of anthropogenic factors on the hydrochemical regime of rivers. In addition, the revealed relationships between the hydrochemical characteristics of water contribute to increasing the automation of water quality monitoring processes.

Key words: ecological forecasting, machine learning, water quality, logistic regression, random forest, pH levels, chloride concentration, environmental impact, data analysis

This study was supported by the National High Level Foreign Experts Introduction Project (G2022014116L). We thank the Introduction of Intelligence Projects for Policy Guidance in Yancheng (YCBX202011) and Scientific Research Project of Yancheng Polytechnic College (ygy2103).

<https://doi.org/10.7546/CRABS.2024.11.06>

Introduction. In the field of managing microecosystems such as forests, rivers, and lakes, human impact is significant. Humans are responsible for living organisms, which requires focusing on preserving flora and fauna while also addressing human health needs. This is especially true for urban aquatic environments, where the primary cause of river pollution is the active growth and development of socio-economic activities along water bodies [1]. It was first established in 1954 that polluted water was the cause of human diseases, with a source of contaminated water identified as the cause of a cholera epidemic in London.

Today, rivers suffer from various pollutants, including household wastewater, agrochemicals, and pesticides, detergents and cleaning agents, household waste and garbage, industrial wastewater, chemical compounds, and oil spills [2, 3]. These sources of pollution significantly alter the chemical composition of water and reduce oxygen levels.

Analytical forecasting plays a crucial role, involving the construction and use of predictive models derived from data. These models are used for forecasting to assist environmental services in decision-making. In everyday life, forecasting refers to potential future events. In analytical forecasting, it refers to assigning values to unknown variables, such as predicting product prices, assessing risk, or classifying situations [4–8]. Papers [4, 5] delve into the relationship between trademark applications and gross domestic product (GDP) using machine learning tools. Subsequently, the study examines trademark filing patterns in individual countries and assesses their correlation with the respective GDP growth rates. RITIKA et al. [6] propose a machine learning model for energy consumption forecasting, utilizing historical data and influencing factors to predict future energy usage patterns accurately. Paper [7] proposes efficient machine learning algorithms for energy pooling market price forecasting, utilizing wavelet transform, fuzzy adaptive PSO, and feed-forward neural networks for accurate predictions. RUPESH et al. [8] utilize the LSTM model for accurate load forecasting in energy management, outperforming XGBoost. The study integrates data processing, model training, and user interaction for tailored forecasts. Machine learning tools are also used in forecasting water quality parameters. In [9] machine learning enables forecasting water quality parameters by modelling complex relationships and detecting anomalies, enhancing the sustainability and security of water resources through real-time monitoring. Paper [10] studies machine learning techniques like Support Vector Machines and Linear Regression forecast Water Quality Index using parameters like pH, turbidity, and conductivity with high accuracy in Tumkur, India. Study [11] compares ARIMA, NNETAR, and Naïve Bayes models for forecasting water parameters like DO, pH, EC, and TDS in freshwater lobster farming, highlighting ARIMA’s efficacy in stream prediction.

The issue of accurate prediction of water quality in urban water environments remains open. Therefore, the goal of this work is to create a high-quality infor-

mation system for environmental forecasting of water quality based on machine learning.

Programme function for environmental forecasting based on machine learning. The main tasks of surface water monitoring are observing, assessing, and forecasting changes in water quality in open water bodies [12,13]. The surface water monitoring system is informational-analytical and does not include management elements. However, it is an essential component of the state environmental management system and quality regulation. These tasks shape monitoring systems, with the block diagram shown in Fig. 1. The diagram illustrates the direct and feedback connections between the main system-forming blocks.

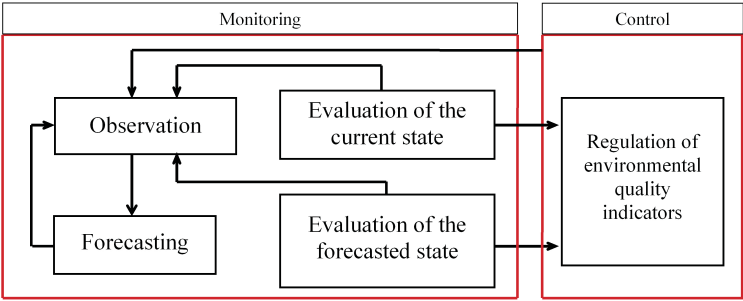


Fig. 1. Structural diagram of the monitoring system

The “Observation” and “State Forecasting” blocks are closely linked, as forecasting the state of the environment is only possible with sufficiently representative information about the actual state (direct connection). On the one hand, forecasting involves understanding the patterns of changes in the natural environment and having the ability for numerical calculations. On the other hand, the focus of the forecast should largely determine the structure and composition of the monitoring network (feedback).

Data characterizing the state of the natural environment, obtained from observations or forecasts, should be evaluated based on the area of human activity in which they are used (using specially selected or developed criteria). Evaluation involves, on the one hand, determining the impact of damage and, on the other hand, selecting optimal conditions for human activity. The results of the assessment of the existing and projected state of the biosphere, in turn, provide an opportunity to clarify the requirements for the observation system. The forecasting process, in turn, can be carried out using machine learning [14,15] with logistic regression and random forest methods.

Logistic regression. Logistic regression is a type of multiple regression that analyzes the relationship between several independent variables (also called regressors or predictors) and a dependent variable [14,16]. Binary logistic regression is used when the dependent variable is binary (i.e., can take only two values).

Logistic regression can be used to estimate the probability that an event will occur for a given subject (e.g., sick/healthy, loan repayment/default, etc.).

All regression models can be expressed in the form of a formula:

$$y = F(x_1, x_2, \dots, x_n).$$

In multiple linear regression, it is assumed that the dependent variable is a linear function of the independent variables, i.e.:

$$\ln(p/(1-p)) = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n,$$

where p is the dependent variable; $x_1, x_2, \dots, x_n$ are independent variables; $b_0, b_1, \dots, b_n$ are coefficients.

To model the probability of a binary dependent variable, a special monotonically increasing function is selected, which can take values in the range from 0 to 1. In binary choice models, the following special functions are typically used: the logistic function; and the standard normal distribution function. Models of binary choice based on the logistic function are called logistic regression or logit models. Models based on the standard normal distribution function are called probit models.

Random forest. Random Forest is a machine-learning algorithm that uses an ensemble of decision trees. The algorithm combines two main ideas: the bagging method by Breiman and the random subspace method proposed by Tin Kam Ho. It is used for classification, regression, and clustering tasks. The main idea is to use a large ensemble of decision trees, each of which, on its own, provides very low-quality classification. However, due to the large number of trees, the overall result is good [14, 17].

The Random Forest method is based on the decision tree method. A Random Forest is a collection of decision trees, and the class of the object being classified is determined by majority voting. A random sample S of size l is generated from the original training sample $D = \{x_i y_i\}_{i=1}^l$.

For the sample S , induce a non-pruned decision tree T_i with a minimum number of observations in terminal nodes equal to $n_{\min}$, recursively following the next sub-algorithm: 1) randomly select p features from the original set of n features, 2) choose the feature among the p features that provide the best decision, 3) split the sample corresponding to the current node into two subsets.

As a result, we obtain an ensemble of decision trees $\{T_i\}_{i=1}^B$. Classification of new observations is performed as follows: let $\hat{y}_l(x) \in \{y_1, \dots, y_l\}$ be a class predicted by the decision tree T_i , that is $T_i(x) = \hat{y}_l(x)$, then $\hat{y}_{rf}^B(x)$ are the class that occurs most frequently in the set $\{\hat{y}_b(x)\}_{b=1}^B$.

Implementation computer system for forecasting water quality parameters. The ecological forecasting software function based on machine learning using the Python programming language was implemented on a local computer

[18–20]. Two different machine learning models, linear regression and random forest regression, were used to predict parameters of the aquatic environment based on a large amount of diverse water environment data. The dataset used for analysis contained hundreds of values over several years, but for this study, only data from 2012 to 2020 were utilized. The goal was to determine whether there is a correlation between various parameters and the increase in key water quality indicators, and if so, which model provides better prediction accuracy. In this case, machine learning models, specifically developed using Scikit-Learn and NumPy,



Fig. 2. Machine learning algorithm for predicting aquatic environment parameters

were used to forecast a specific parameter based on data collected by a computer system for monitoring.

In Fig. 2, we assess the accuracy of our predicted indices compared to real data. To achieve this, we followed several steps outlined in Fig. 2.

The algorithm begins with the “Reading CSV data” step, where we read `gdp_services_percentage.csv`, `gdp_growth_data.csv`, and `trademarks_filed_count.csv` into pandas DataFrames. The next step is “Data preparation”, involving the extraction and organization of data needed for analysis, such as the number of parameters and predictions for specific parameters. For the aquatic environment, we used relevant data from 2012 to 2020 and created feature matrix X and target vector y . The program then moves to “Model training”, where we train two machine learning models, linear regression and random forest, on the prepared data. After training, we display the results with “Result plotting”, visualizing the actual and predicted values for each model to better understand their effectiveness.

Next, we perform forecasting using the trained models to predict future parameter values. Finally, we use “Structuring and saving predictions”, storing the forecasts in a pandas DataFrame and saving them to a CSV file for further analysis.

Measured and calculated data in the urban aquatic environment study are consolidated into a database stored on a server.

Results and discussions. The results of forecasting water quality parameters of the Southern Bug river are presented in Fig. 3. Analyzing the obtained results (Table 1), we can see that the machine learning models were able to predict “pH” and “Chlorides” with high accuracy.

T a b l e 1
Results of the forecast of water quality parameters

Parameter	Values by linear regression	Values by random forest	Real value	Error value for linear regression	Error value for random forest
Chloride	43.2	58	62.88	31.298	7.760
pH	9.65	8.45	8.19	17.827	0.672
General hardness	4.75	6.42	6.65	28.571	3.459
Dissolved oxygen	6.2	11.4	10.71	42.111	6.443
Nitrates	4.8	2.8	1.94	147.42	44.33
Nitrite	0.069	0.034	0.02	245	70

The linear regression model demonstrates impressive accuracy with a prediction error of about 20 units for chloride and 1.4 units for the pH value, indicating its ability to reliably predict indicator values in complex multidimensional data.

The pH prediction by the random forest model shows minor deviations from the actual average values for 2021: 5 units for chloride and 0.05 units for pH.



Fig. 3. Prediction using machine learning: a, c, e, g – linear regression; b, d, f, h – random forest

These deviations within the norm indicate the reliability of the models in the context of real operating conditions. Both models showed good consistency in their predictions, with relatively small deviations from the actual values.

From a scientific point of view, it is useful to search for and evaluate the closeness of relationships between the content of individual hydrochemical characteristics in the water. This allows for selecting analytical calculation dependencies, conducting operational assessments and forecasting water quality, and automating the calculations of dissolved chemical substances flow. These relationships do not always exist, depending on the complexity of the conditions forming the chemical composition of river waters in catchments, and the level of anthropogenic load impact, which together can alter the natural hydrochemical regime of rivers.

In this study, visualizations created in Jupyter Notebook play a crucial role in illustrating the process of creating machine learning models using two algorithms: linear regression and random forest regression. These results provide a better understanding of the behaviour and effectiveness of the models, offering insights into the complex relationships between all water environment parameters and the forecast of a specific selected parameter.

Conclusions. The article investigates the application of machine learning techniques for ecological forecasting and water quality analysis in the Southern Bug river. By employing binary logistic regression and random forest regression models, the study evaluates the impact of various environmental parameters on key water quality indicators, such as pH levels, chloride, general hardness, dissolved oxygen, and nitrates nitrite.

The research involved preparing and processing data from 2012 to 2020, with both models demonstrating effective performance in predicting water quality, as evidenced by their accuracy. The findings highlight that wastewater discharge remains a significant pollutant affecting rivers. The study's results contribute to the development of effective water resource management strategies, improve predictions of the ecological status of water bodies, and enhance the assessment of anthropogenic impacts on river hydrochemical regimes. Furthermore, the established relationships between water quality parameters support real-time monitoring and the automation of dissolved chemical substance flow calculations, promoting more informed and timely environmental management decisions.

REFERENCES

- [1] DAVIES-COLLEY R. J., D. G. SMITH, R. C. WARD et al. (2011) Twenty Years of New Zealand's National Rivers Water Quality Network: Benefits of Careful Design and Consistent Operation, JAWRA Journal of the American Water Resources Association, **47**(4), 750–771, <https://doi.org/10.1111/j.1752-1688.2011.00554.x>.

- [2] FÖLSTER J., R. JOHNSON, M. FUTTER, A. WILANDER (2014) The Swedish Monitoring of Surface Waters: 50 Years of Adaptive Monitoring, *AMBIO*, **43**(S1), 3–18, <https://doi.org/10.1007/s13280-014-0558-z>.
- [3] TAUFİK I. S., T. HAIYUNNISA, H. S. ALAM (2016) Design and Implementation of Water Quality Monitoring for Eel Fish Aquaculture, 2016 International Symposium on Electronics and Smart Devices (ISESD), <https://doi.org/10.1109/ISESD.2016.7886718>.
- [4] ROBOTKO A., A. TOPALOV, V. LUKASHOVA et al. (2023) A Machine Learning Study on World Economic Impact of Intellectual Property Rights on GDP Growth with a Focus on Poland, 2023 13th International Conference on Advanced Computer Information Technologies (ACIT), Wrocław, Poland, 261–264, <https://doi.org/10.1109/ACIT58437.2023.10275475>, <https://ieeexplore.ieee.org/document/10275475>.
- [5] ROBOTKO S., A. TOPALOV, S. NEKRASOV et al. (2023) Machine Learning and Modeling of the Impact of Trademark Filings on GDP Growth based on Python, MoM-LeT+DS 2023: 5th International Workshop on Modern Machine Learning Technologies and Data Science, June 3, 2023, Lviv, 65–76, <https://ceur-ws.org/Vol-3426/paper6.pdf>.
- [6] RITIKA V., R. GUPTA, A. SAHU, N. GOYAL (2024) Study on energy consumption forecasting using machine learning, *International Journal of Innovative Research in Computer Science & Technology*, <https://doi.org/10.55524/csistw.2024.12.1.68>.
- [7] SIVANANDA REDDY Y., B. SRILATHA, Y. CHALAMALASETTI, P. DILIP (2024) Energy Pooling Market Price Forecasting using Machine Learning, *International Journal for Multidisciplinary Research*, <https://doi.org/10.36948/ijfmr.2024.v06i03.19133>.
- [8] RUPESH K. C., R. RENUSHREE, S. DEEPIKA et al. (2024) Load Forecasting using Machine Learning Model, <https://doi.org/10.1109/icsses62373.2024.10561429>.
- [9] SHASHANK Y., S. AMITH, S. SAHANA, A. S. SASTRY (2024) Monitoring of Water Quality using Machine Learning – A Review, *International Journal for Research in Applied Science and Engineering Technology*, <https://doi.org/10.22214/ijraset.2024.61578>.
- [10] ARUN KUMAR T., D. THIMALAPUR, B. EARAPPA et al. (2024) Machine Learning for Water Quality Index Forecasting, <https://doi.org/10.46604/emsi.2024.12870>.
- [11] ABDELMONEIM A. B., N. S. MOHD SABLI, M. F. JAMLOS et al. (2024) IoT-based Machine Learning Comparative Models of Stream Water Parameters Forecasting for Freshwater Lobster, *Journal of Advanced Research in Applied Mechanics*, <https://doi.org/10.37934/aram.117.1.137149>.
- [12] MUKTA M., S. ISLAM, S. D. BARMAN, A. W. REZA (2019) IoT based Smart Water Quality Monitoring System, 2019 IEEE 4th International Conference on Computer and Communication Systems, <https://doi.org/10.1109/CCOMS.2019.8821656>.
- [13] BEHMEI S., M. DAMOUR, R. LUDWIG, M. J. RODRIGUEZ (2019) Optimization of River and Lake Monitoring Programs Using a Participative Approach and an Intelligent Decision-Support System, *Applied Sciences*, **9**(19), 4157, <https://doi.org/10.3390/app9194157>.

- [14] KELLEHER J. D., B. M. NAMEE, A. D'ARCY (2015) *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, The MIT Press, 624 pp.
- [15] HUYEN C. (2022) *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*, O'Reilly Media, 1st edition, 386 pp.
- [16] RYMARCZYK T., E. KOZŁOWSKI, G. KŁOSOWSKI, K. NIDERLA (2019) Logistic regression for machine learning in process tomography, *Sensors*, **19**(15), 3400.
- [17] LIU Y., Y. WANG, J. ZHANG (2012) New machine learning algorithm: Random forest, *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14–16, 2012, Proceedings 3*, Springer Berlin Heidelberg.
- [18] MUELLER J., L. MASSARON (2019) *Python for Data Science, For Dummies*, 2nd ed.
- [19] VOTHIHONG P., M. CZYGAN, I. IDRIS et al. (2017) *Python: End-to-end Data Analysis*, Packt.
- [20] MCKINNEY W. (2022) *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*, O'Reilly Media, 3rd edition, 579 pp.

¹*Department of Science and Technology, Yancheng Polytechnic College, Yancheng, China*
e-mail: and_bsb@126.com

²*Scientific Department, Yunzhou (Yancheng) Innovation Technology Co., Ltd, China*
e-mail: jianjun.wang@yunzhou-tech.com

³*Scientific Department, International Academy of Marine Sciences, Technologies and Innovations, 54025 Mykolaiv, Ukraine*
e-mail: sergiy.ryzhkov@me.com

⁴*Laboratory of Ecology, Institute of Water Problems and Land Reclamation, 03022 Kyiv, Ukraine*
e-mail: Water_2019@ukr.net

⁵*Department of Computerized Control Systems, Admiral Makarov National University of Shipbuilding, 54025 Mykolaiv, Ukraine*
e-mails: topalov_ua@ukr.net, zivenkoav@zivenko.com.ua

⁶*Department of Computer Systems and Networks, Ivano-Frankivsk National Technical University of Oil and Gas, 76019 Ivano-Frankivsk, Ukraine*
e-mails: serhii.babchuk@nung.edu.ua, taras.harasymiv@nung.edu.ua