A WATER QUALITY PREDICTION MODEL FOR LARGE-SCALE RIVERS BASED ON PROJECTION PURSUIT REGRESSION IN THE YANGTZE RIVER

by

Ze-Ji YI[a], Xiao-Hua YANG[a*], and Yu-Qi LI[b]

[a] State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing, China
[b] Department of Urban Design and Planning, University of Washington, Seattle, Wash., USA

Original scientific paper
https://doi.org/10.2298/TSCI2203561Y

In recent decades, the Yangtze River Basin, which carries hundreds of millions of people and a substantial economic scale, has been plagued by water quality deterioration, threatening considerably sustainable development. In this paper, a sample set is established based on the water quality indexes of chemical oxygen demand and dissolved oxygen obtained by week-by-week monitoring on the main stream of the Yangtze River in Panzhihua, Yueyang, Jiujiang, and Nanjing from 2006 to 2018. The twelve characteristic variables are selected by random forest technique, and the week-by-week dynamic prediction models of chemical oxygen demand and dissolved oxygen at each section of main stream are established by the projection pursuit regression, which can effectively predict the water quality dynamics of the Yangtze River main stream.

Key words: *water quality, dynamic prediction model, random forest, projection pursuit regression, Yangtze River*

Introduction

Water resources are primary natural resources and strategic economic resources, which are irreplaceable essential elements of human survival and social development. Good water quality is of great significance to realize the sustainable development of society [1, 2]. With economic development, population expansion, and the decline of river self-purification capacity, the Yangtze River Basin, which carries hundreds of millions of people and a large economic scale, has also suffered varying degrees of pollution [3]. Water pollution reduces the function of water use and intensifies the shortage of water resources, which is a threat to sustainable development [4].

In order to solve this problem effectively, reasonable planning and integrated management of water resources in the Yangtze River Basin is imperative [5]. The reliable and accurate prediction of water quality dynamics is an important basis for maintaining and improving the water quality of the Yangtze River.

Water quality prediction is to establish the corresponding mapping relationship between the multivariate monitoring data and the change of water quality parameters [6]. The methods can be classified into two categories. One is to use mathematical statistics or statisti-

* Corresponding author, e-mail: xiaohuayang@bnu.edu.cn

cal learning methods combined with water quality series data to predict. The other method is to use a water quality model to predict. The advantage of the former is that the method is simple and the required parameters are less, and the disadvantage is that a long series of measured water quality data is needed, and sometimes missing data distribution is not known [7] or the data are stochastically uncertain [8, 9]. The latter method is more complex and requires more parameters, but it is theoretical [10]. In practice, however, the first method is more applied in water quality prediction because of fewer demand parameters. At present, the most representative methods are machine learning, such as the neural network method, swarm intelligence algorithm, *etc.* [11, 12]. These methods have strong non-linear mapping ability, learning ability, and fault tolerance.

Projection pursuit is a kind of statistical method to process and analyze high-dimensional data. Its basic idea is to project high-dimensional data onto low-dimensional (1~3 dimensional) subspaces to find a projection that reflects the structure or features of the original high-dimensional data in order to achieve the purpose of research and analysis of high-dimensional data [13]. At present, projection pursuit regression is less used in prediction. According to previous attempts, the effect of projection pursuit regression in water quality prediction is often better than that of statistical learning methods such as back propagation neural network. Therefore, this study attempts to apply it to the prediction of water quality in the Yangtze River main stream.

The chemical oxygen demand (COD) and dissolved oxygen (DO) are the most comprehensive and effective indicators to reflect the water quality [14]. Therefore, this study simulates the COD and DO dynamics of the four representative national monitoring sections on the Yangtze River main stream by projection pursuit regression, which is helpful to evaluate the applicability of projection pursuit regression in river water quality prediction. Combined with random forest to screen characteristic variables, the input of projection pursuit regression prediction model is adjusted in order to predict the water quality change of the Yangtze River main stream more accurately.

Data and methods

Data description

This study collected the COD and DO values of the key national water quality monitoring sections on the main stream of the Yangtze River from 2006 to 2018, and Panzhihua, Yueyang, Jiujiang, and Nanjing were selected to compare the reliability of water quality prediction models, which are arranged as fig. 1.

The data were obtained from the automatic water quality monitoring weekly reports published on China National Environmental Monitoring Centre, and some of them were from the Economic Forum of RUC.

Projection pursuit regression and random forest

Projection pursuit regression is proposed by Peter Hall in its basic form:

$$E(Y \mid X_1, X_2, \ldots, X_p) = \mu_y + \sum_{m=1}^{M} \beta_m \varphi_m (a_m^T x) \tag{1}$$

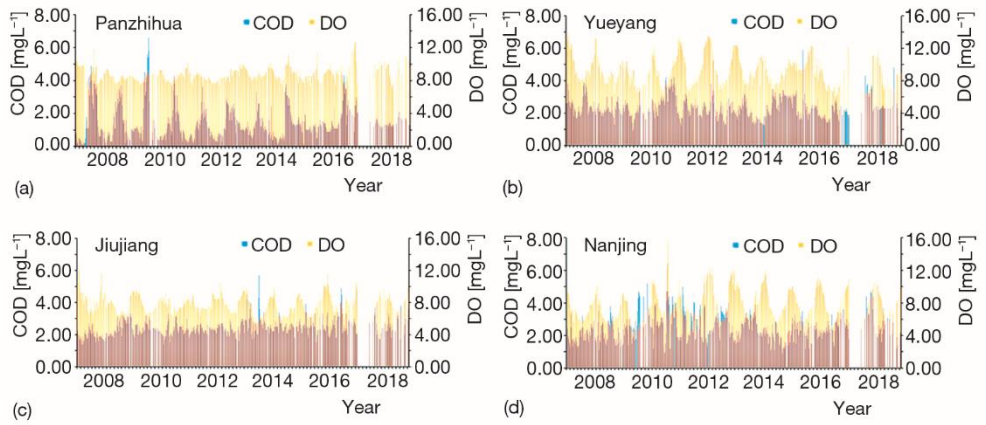

Figure 1. Water quality dynamics of the main stream of the Yangtze River during 2006-2018

Projection pursuit regression modeling is to choose the very β_m, φ_m and a_m with a minimized objective function:

$$E\left\{\left[y-\mu_y-\sum_{m=1}^{M}\beta_m\varphi_m(a_m^T x)\right]^2\right\} \tag{2}$$

During the analysis, another group of samples that did not participate in the training can be applied to test the reliability of the model. According to the projection pursuit regression model established by the training sample, the average relative error between the simulated predicted value and the actual observed value can be calculated by mean absolute percentage error (MAPE).

Random forest is a classification and prediction model proposed by Breiman [15]. The new training sample set is generated by repeated random sampling of *n* samples from the original training sample set *N* by bootstrap method, and then *k* classification trees are generated according to the bootstrap sample set to form the random forest.

Random forests can estimate the importance of variables by random sampling, so they can be used to screen characteristic variables so that the overfitting caused by the relative shortage of sample size can be effectively avoided in the training of subsequent prediction models [16].

Methodology

The COD and DO dynamics of each water quality monitoring section are found to be periodic, among which the annual cycle is the most significant [17]. Therefore, we selected the COD and DO of all 24 weeks before the prediction period as the feature variables and selected some variables from the 48 feature variables as the input variables of the prediction model through the variable importance estimation method of random forest. Some of the importance ranking results of characteristic variables are shown in tab. 1. Finally, we use the same water quality indicators in the previous 12 weeks to predict the COD and DO in the current week.

Table 1. Importance ranking of feature variables

DO		COD		DO		COD	
Degree of importance	Characteristic variable	Degree of importance	Characteristic variable	Degree of importance	Characteristic variable	Degree of importance	Characteristic variable
0.328	X48	0.274	X24	0.117	X37	0.107	X14
0.248	X47	0.220	X23	0.113	X9	0.106	X12
0.215	X46	0.211	X22	0.111	X24	0.105	X37
0.195	X45	0.169	X21	0.109	X23	0.105	X17
0.183	X44	0.141	X20	0.107	X17	0.105	X16
0.145	X43	0.121	X18	0.105	X8	0.103	X36
0.130	X42	0.120	X19	0.104	X38	0.102	X10
0.120	X41	0.112	X13	0.104	X4	0.102	X33
0.118	X36	0.107	X15	0.103	X40	0.101	X30

This study is expected to establish a dynamic prediction model of the COD and DO of the four key national water quality control sections on the main streams of the Yangtze River in Panzhihua, Yueyang, Jiujiang, and Nanjing, so as to predict the water quality of the Yangtze River at the sections next week to provide key data input for the water quality warning system of the main stream.

Results and discussion

The simulation of the training set

On the basis of incomplete weekly water quality data from 2006 to 2018, a total of 484 samples meet the requirements of the prediction model. The last 48 weeks of continuous water quality data were selected as the verification set or prediction period, and the other 436 samples formed a training set. The COD and DO dynamics of main stream water quality section were fitted by projection pursuit regression. The simulation effect of projection pursuit regression on water quality dynamics is shown in fig. 2.

Judging from the fitting curve, the water quality prediction models established at the four sections can effectively simulate the water quality dynamics, but it is obvious that the fitting accuracy of projection pursuit regression to the DO dynamics is higher than that of the COD dynamics simulation. In order to measure the reliability of the model, we characterize the water quality simulation accuracy by calculating the coefficient of determination and relative error of the proposed sample, as shown in tab. 2.

Table 2. Coefficient of determination and relative error of the proposed samples

Water quality monitoring section	COD		DO	
	R	MAPE	*R*	MAPE
Panzhihua	0.869	9.69%	0.835	2.41%
Yueyang	0.731	9.91%	0.924	3.87%
Jiujiang	0.635	8.39%	0.916	3.07%
Nanjing	0.694	14.92%	0.931	4.71%

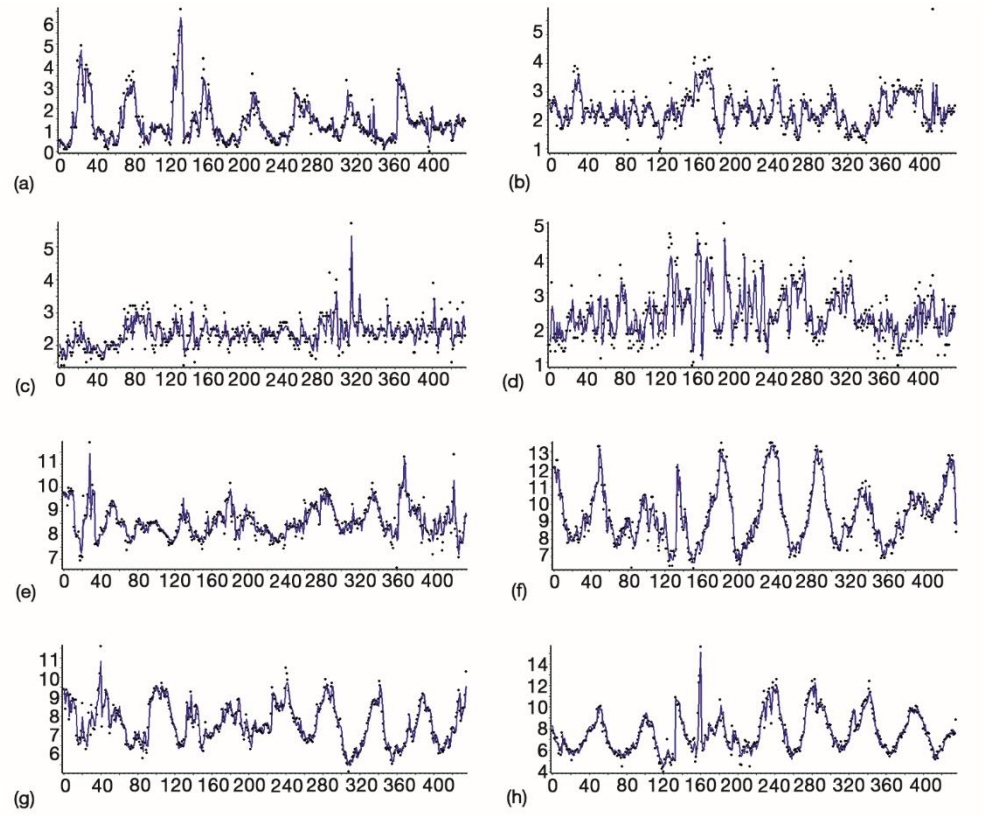

Figure 2. The simulation effects of projection pursuit regression on water quality dynamics; COD and DO dynamics of the national water quality monitoring sections on the main stream of the Yangtze River in Panzhihua (a, e), Yueyang (b, f), Jiujiang (c, g), and Nanjing(d, h)

Table 2 evaluates the reliability of the water quality prediction model based on projection pursuit regression from 2-D of coefficient of determination and relative error and confirms the above judgment. The coefficient of determination of the prediction model for the main stream COD is relatively lower, and the relative error of that is relatively higher. The results of the main stream water quality simulation at the four key monitoring water quality sections support this conclusion. According to commonly used standards, the MAPE of the main stream COD simulation at the four water quality sections are more than 5%, while the MAPE of the main stream DO simulation are less than 5%, so only the prediction models for DO can be considered reliable.

To explain its reasons, the types and quantities of data needed for COD prediction at a section of the Yangtze River main stream in the current week are large, and the characteristic variables of the water quality prediction model have very limited ability to explain the explanatory variables. According to the laws of pollutant migration and transformation in the river water body, for the weekly COD prediction, the ability of the relevant data of pollutants in the upstream water body to explain the prediction variables should be greater than the COD data in the early stage at the river sections. Moreover, compared with DO prediction, the uncertain effects of human activities such as sewage discharge in COD prediction are more direct, and these effects are rarely reflected in the characteristic variables of the model.

The simulation of the validation set

The DO of the continuous 48 week verification set selected in this paper was predicted according to the trained prediction models of the main stream DO. The test results of the predicted values compared with the actual values of the prediction period are shown in fig. 3.

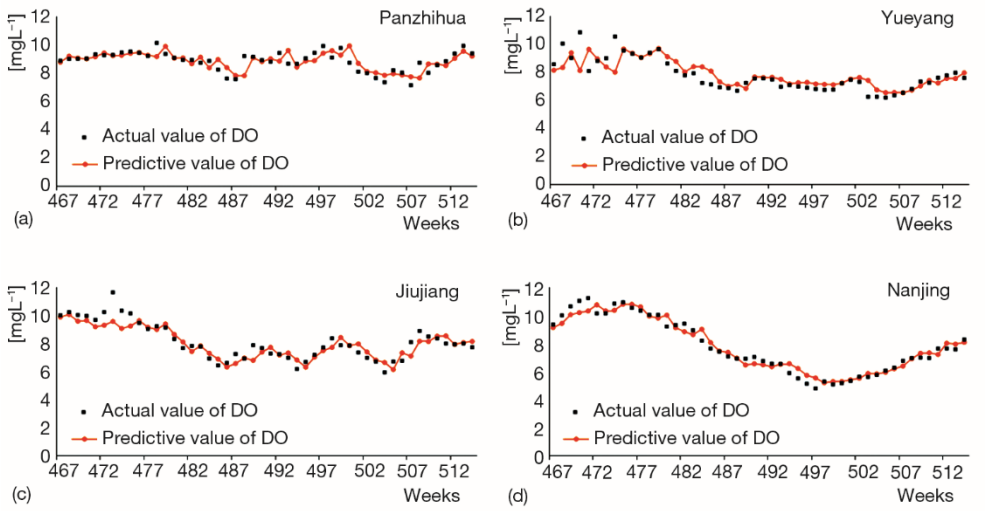

Figure 3. Simulation results of DO prediction models during validation period

According to the previous verification results of model prediction, the model can meet the weekly forecast demand of the Yangtze River main stream DO and contribute to the early warning of the main stream water quality. Compared to COD, the dynamics of the DO are more closely related to the ecological hydraulic characteristics of a particular river reach, so the characteristic variables containing the relevant characteristic information of the river reach have a stronger ability to explain the prediction variables. This may be an important reason for the relative reliability of the main stream DO prediction model. Furthermore, significant annual cycles of dissolved oxygen dynamics in surface rivers also improve the accuracy of DO prediction.

Conclusions

The week-by-week dynamic prediction models of COD and DO at four national water quality monitoring sections on the main stream of the Yangtze River were established by projection pursuit regression based on the characteristic variables selected by random forest. According to the aforementioned research, projection pursuit regression can effectively predict the water quality dynamics of the Yangtze River main stream. Compared with the COD dynamics, the DO dynamics in the early stage of the same sections contain more information that can effectively explain the current DO at the water quality sections, so the main stream DO prediction model is better fitted and the prediction accuracy is higher. A reliable prediction model for DO dynamics is constructed, and the output results can provide effective support for the Yangtze River water quality warning system.

Acknowledgment

This work was supported by the National Key Research Program of China (No. 2017YFC0506603), the State Key Program of National Natural Science of China (No. 41530635), and the Project of the National Natural Foundation of China (No. 52179001).

References

[1] Patterson, J. J., *et al.*, Understanding Enabling Capacities for Managing the 'Wicked Problem' of Non-point Source Water Pollution In Catchments: A Conceptual Framework, *Journal of Environmental Management*, *128* (2013), Oct., pp. 441-452
[2] Xue, Q. R., *et al.*, A Three-Stage Hybrid Model for the Regional Assessment, Spatial Pattern Analysis and Source Apportionment of the Land Resources Comprehensive Supporting Capacity in the Yangtze River Delta Urban Agglomeration, *Science of the Total Environment*, *711* (2020), Apr., pp. 1-18
[3] Jiang, Y., China's Water Scarcity, *Journal of Environmental Management*, *90* (2009), 11, pp. 3185-3196
[4] Yang, X. H., *et al.*, Hierarchy Evaluation of Water Resources Vulnerability under Climate Change in Beijing, China, *Natural Hazards*, *84* (2016), 1, pp. 63-76
[5] Sun, B. Y., *et al.*, Evaluation of Water Use Efficiency of 31 Provinces and Municipalities in China Using Multi-Level Entropy Weight Method Synthesized Indexes and Data Envelopment Analysis, *Sustainability*, *11* (2019), 17, pp. 1-8
[6] Emamgholizadeh, S., *et al.*, Prediction of Water Quality Parameters of Karoon River (Iran) by Artificial Intelligence-Based Models, *International Journal of Environmental Science and Technology*, *11* (2014), 3, pp. 645-656
[7] Zhang, Z. W., *et al.*, Evidence Integration Credal Classification Algorithm *vs.* Missing Data Distributions, *Information Sciences*, *569* (2021), Aug., pp. 39-54
[8] Li, L., *et al.*, Modelling and Filtering for a Stochastic Uncertain System in a Complex Scenario, *Thermal Science*, *25* (2021), 2, pp. 1411-1424
[9] Liu, L., *et al.*, Distributed State Estimation for Dynamic Positioning Systems with Uncertain Disturbances and Transmission Time Delays, *Complexity*, *2020* (2020), ID 7698504
[10] Memon, F. A., *et al.*, Assessment of Gully Pot Management Strategies for Runoff Quality Control Using a Dynamic Model, *Science of the Total Environment*, *295* (2002), 1-3, pp. 115-129
[11] Yang, X. H., *et al.*, A Fractional-Order Genetic Algorithm (FOGA) for Parameter Optimization of the Moisture Movement in A Bio-Retention System, *Thermal Science*, *23* (2019), 4, pp. 2343-2350
[12] Yang, X. H., *et al.*, Comprehensive Assessment for Removing Multiple Pollutants, by Plants in Bioretention Systems, *Chinese Science Bulletin*, *59* (2014), 13, pp. 1446-1453
[13] Zhao, J., *et al.*, Dynamic Risk Assessment Model for Water Quality on Projection Pursuit Cluster, *Hydrology Research*, *43* (2012), 6, pp. 798-807
[14] Raheli, B., *et al.*, Uncertainty Assessment of the Multilayer Perceptron (MLP) Neural Network Model with Implementation of the Novel Hybrid MLP-FFA Method for Prediction of Biochemical Oxygen Demand and Dissolved Oxygen: a Case Study of Langat River, *Environmental Earth Sciences*, *76* (2017), 14, pp. 502-517
[15] Breiman, L., Random Forests, *Machine Learning*, *45* (2001), 1, pp. 5-32
[16] Kim, S., *et al.*, Assessing the Biochemical Oxygen Demand Using Neural Networks and Ensemble Tree Approaches in South Korea, *Journal of Environmental Management*, *270* (2020), ID 110834
[17] Huang, H., *et al.*, Identification of River Water Pollution Characteristics Based on Projection Pursuit and Factor Analysis, *Environmental Earth Sciences*, *72* (2014), 9, pp. 3409-3417

Paper submitted: March 1, 2020
Paper revised: August 10, 2021
Paper accepted: August 10, 2021