Research on modeling method of continuous spectrum water quality online detection based on random forest*

LI Wen, HAO Sijia, ZHOU Hao, and LIU Ying**

Institute of Mechanical and Electrical Engineering, North China University of Technology, Beijing 100144, China

(Received 18 July 2022; Revised 23 September 2022)

©Tianjin University of Technology 2023

It's common to use the method of continuous spectroscopy in water quality testing. But there're some problems with it. For example, the scanning results have a large number of nonlinear signals, and the covariance between variables is serious, which can lead to a decrease in the model prediction accuracy. In this paper, the standard solutions of nitrate nitrogen (NO_3-N) and nitrite nitrogen (NO_2-N) were used as the subject to be tested, and the data of the scanned waves and absorbance were obtained by use of spectral detector. The data were processed by noise reduction first and then the random forest (RF) algorithm was adopted to establish the regression relationship between concentration and absorbance. For comparison, partial least squares (PLS) and support vector machine (SVM) algorithm models were also established. For the same given data, the three reverse models can make the projection of the concentration respectively. The experimental results show that the RF algorithm predicts NO_2-N concentrations significantly better than the SVM algorithm and PLS algorithm. This proves that the RF algorithm has good prediction ability in spectral water quality detection because of its high model accuracy and better adaptability, which could be a reference for similar research on continuous spectral water quality online detection.

Document code: A **Article ID:** 1673-1905(2023)02-0095-6

DOI https://doi.org/10.1007/s11801-023-2127-9

The water quality evaluation is mainly manifested in the comprehensive factors such as multi-chemical element pollution. At present, spectral analysis method has been widely studied and applied in the field of water quality detection[1]. However, the online detection of water quality has the following problems[2]. It cannot respond to changes in water quality in advance. The amount of raw spectral data of the scanned nonlinear signals is huge. The high correlation of data variables will lead to multicollinearity. Therefore, it is necessary to choose an appropriate modeling method for prediction. At present, the commonly used methods for spectral modeling mainly include partial least squares (PLS), support vector machine (SVM) and random forest (RF). The model established by the PLS[3,4] method is widely used and can solve the multicollinearity problem between independent variables well, but there is a large generalization error. The SVM algorithm needs to use the interactive verification method to estimate the penalty factor C and the kernel function parameter g. In order to ensure the correctness of the data in the overall situation, the SVM algorithm will add the data fault tolerance rate, which will increase the amount of calculation[5]. The RF[6] algorithm can identify complex nonlinear relationships between independent variables and response variables. The RF algorithm has strong analytical ability to nonlinear data, and is suitable for nonlinear signals scanned by spectrometers. The algorithm has a higher accuracy and performs well in classification and regression. RF algorithms have broad application prospects in the field of machine learning[7]. LI et al[8] proposed an early warning method for sudden water pollution incidents using an RF model. MOHAMMADI et al[9] investigated predictive modeling algorithms, namely RF, for classification of nanofluid solutions based on viscosity values in the presence of different concentrations of salinity, silica nanoparticles, and polyacrylamide solutions. BARAKA et al[10] tested groundwater and proposed that RF is the best model for predicting the pollution of groundwater fluoride in the study area.

In this research, the band-absorbance data were denoised by spectral scanning of standard solutions of NO_3-N and NO_2-N with different concentrations. The regression of concentration and absorbance for the parameters to be measured is realized by applying RF modeling method. The model was used to predict the deviation of the concentration and absorbance of the parameters to be measured in the water samples from the actual values, which is the RF model for NO_3-N and NO_2-N (NO_3-N-RF, NO_2-N-RF). The accuracy of the RF algorithm in

* This work has been supported by the National Natural Science Foundation of China (No.51205005), and the Beijing Science and Technology Innovation Service Ability Building (No.PXM2017-014212-000013).

** E-mail: hsj991108@163.com

estimating NO_3-N and NO_2-N concentrations was also tested by comparing the R^2 and root-mean-square error (*RMSE*) of the three models with PLS and SVM models (NO_3-N-PLS, NO_2-N-PLS, NO_3-N-SVM, NO_2-N-SVM) with the same parameters. This model can not only solve the current problems of online water quality detection, but also has important significance to enhance the multi-dimensional and comprehensive understanding of water environment quality, and provide reference for continuous spectrum water quality detection research.

RF is an integrated classifier based on decision tree algorithm, which improves prediction accuracy and predicts samples by random feature selection[11], and it has been described as a method representing the state of the art in integrated learning[12]. The RF algorithm becomes more robust and more accurate as the number of decision trees increases in model learning with sample variables. The RF algorithm is less prone to overfitting, can highly explain the importance of each feature, and can guarantee the accuracy of the data even if it is partially missing. The RF algorithm is more capable of resolving nonlinear data[13], and in water quality detection the signal scanned by the spectrometer is nonlinear, the amount of raw spectral data is cumbersome, and there are multiple covariance problems between variables, and for the spectrally sensitive band of the measured solution, the spectral redundancy at the absorption peak can be reduced to improve the model accuracy .

The specific forecasting process can be divided into the following four steps.

Assume the number of cases in the training set is *N*. Then, sample of these *N* cases is taken at random but with replacement.

If there are *M* input variables or features, a number $m_{\text{try}}<M$ is specified such that at each node, *m* variables are selected at random out of the *M*. The best split on these *m* is used to split the node. The value of *m* is held constant while we grow the forest.

Each tree is grown to the largest extent possible and there is no pruning.

After completing the above three steps, the algorithm will finally get a decision tree and repeat *K* times to get N_{tree} decision trees (the number of trees N_{tree} is *K*).

For the regression problem, the RF algorithm can predict the unknown parameters through these N_{tree} decision trees, take the average of the prediction results as the output[14], and the specific process is shown in Fig.1.

The m_{try} and N_{tree} are two important parameters in the modeling process. m_{try} is the number of *n* variables randomly selected from *p* variables in the construction of the decision tree as the number of branching nodes of the decision tree. The m_{try} parameters are set from 1 to *n* by traversal, *n* trials are performed and the error rate of each trial is output, and the value with the lowest error rate is selected as the optimal m_{try} value.

Modeling is performed by the optimal number of variables m_{try} and the correspondence between the model error rate and the number of decision trees is output, the value of the number of decision trees after the model error rate is stabilized is used as the N_{tree} value for constructing the RF regression model. The most appropriate number of decision trees N_{tree} and the number of features m_{try} in the regression model is chosen by the adaptive function in the “RF” package of the R language called.

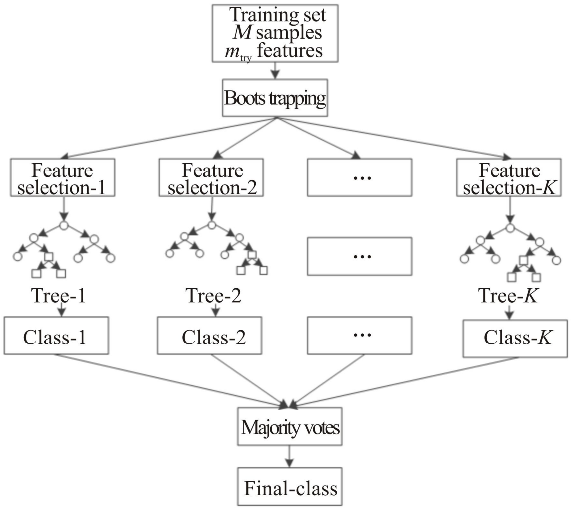

Fig.1 Flow chart of RF

The instruments and equipment used in the experiment are pulse xenon light source, micro-spectrometer, optical fiber, quartz cuvette, volumetric flask, beaker, graduated cylinder, plastic tip dropper, single-channel adjustable volume pipette for research plus 0.1 μL to 2 μL and 0.5 mL to 5 mL.

The reagents used in the experiment are deionized water with grade EW-I in “GB11446-1-2013 National Standard for Deionized Water”, NO_3-N standard solution numbered “GSB-04-2837-2011” with concentration of $c(NO_3\text{-}N)=100\ \text{mg}\cdot\text{L}^{-1}$, and the NO_2-N standard solution numbered “GSB-04-2840-2011” with concentration of $c(NO_2\text{-}N)=100\ \text{mg}\cdot\text{L}^{-1}$. The specific experimental process is shown in Fig.2.

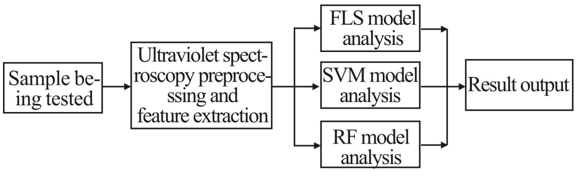

Fig.2 Flow chart of experimental process

The signal scanned by the spectrometer is a “non-linear non-stationary” signal. The analysis and noise reduction processing method of Hilbert-Huang transform (HHT)[15] is more accurate than the traditional signal noise reduction processing method, and HHT is adaptive and does not need to set a basis function, so HHT is selected as the data processing method. The

HHT method was used to perform spectral scanning noise reduction processing on standard solutions with different concentrations of NO_3-N and NO_2-N to provide data for model establishment.

NO_3-N is one of the three nitrogens (nitrate nitrogen, ammonia nitrogen, and total nitrogen) in water, which can reflect the pollution degree of water bodies and is an important indicator for judging eutrophication of water bodies[16,17]. The range of nitrate nitrogen was set to 0—10 mg·L^{-1}. Dissolve 0.2 mL, 1.0 mL, 2.0 mL, 4.0 mL, 6.0 mL, 8.0 mL, 10.0 mL of nitrate nitrogen standard solution into a 100 mL volumetric flask, and add ultrapure water to dilute to the mark. It can be diluted to obtain the working point of nitrate nitrogen standard solution of 0.2 mg·L^{-1}, 1.0 mg·L^{-1}, 2.0 mg·L^{-1}, 4.0 mg·L^{-1}, 6.0 mg·L^{-1}, 8.0 mg·L^{-1}, 10.0 mg·L^{-1}. Taking deionized water as a reference, the spectra of NO_3-N solutions with different concentrations were scanned, and the results after noise reduction are shown in Fig.3. It can be seen that in the wavelength range of 200—300 nm, the trend of nitrate nitrogen spectral curves of solutions with different concentrations is similar. With the increase of solution concentration, the absorption peak gradually increases in the wavelength range of 200—210 nm. In the wavelength range of 210—220 nm, a peak appears due to the highest absorption value of dissolved organic matter in the solution, and the final peak curve tends to be flat at 250 nm.

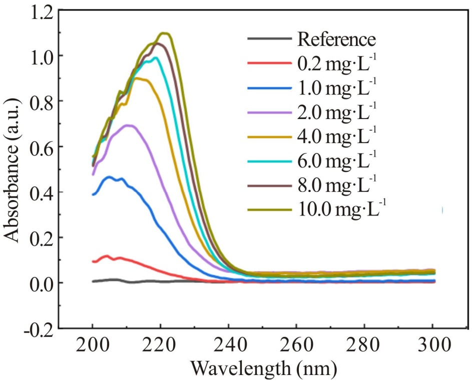

Fig.3 Wavelength-absorption peak curves of NO_3-N solutions

NO_2-N is an intermediate product of nitrogen-containing organic matter in water becoming nitrate. If its content is too high, it means that there is still a risk of pollution[18]. The range of nitrite nitrogen was set to 0—1 mg·L^{-1}. Dissolve 0.2 mL, 0.4 mL, 0.6 mL, 0.8 mL, 1.0 mL of nitrite nitrogen standard solution in a 100 mL volumetric flask, and add ultrapure water to dilute to the marked line. It can be diluted to obtain 0.2 mg·L^{-1}, 0.4 mg·L^{-1}, 0.6 mg·L^{-1}, 0.8 mg·L^{-1}, 1.0 mg·L^{-1} of nitrite nitrogen standard solution working point. Taking deionized water as a reference, the spectra of NO_2-N solutions with different concentrations were scanned, and the results after noise reduction are shown in Fig.4. It can be seen that in the wavelength range of 200—300 nm, the trend of nitrate nitrogen spectral curves of solutions with different concentrations is similar. With the increase of solution concentration, the absorption peak gradually increases in the wavelength range of 200—210 nm. In the band around 210 nm, a peak appears due to the highest absorption value of dissolved organic matter in the solution, and the final peak curve tends to be flat at 250 nm.

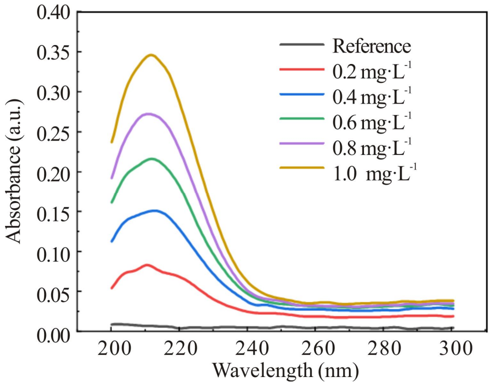

Fig.4 Wavelength-absorption peak curves of NO_2-N solutions

The PLS and SVM models were selected for comparison, and the prediction accuracy of the RF model for the concentration-absorbance of nitrate nitrogen solution and nitrite nitrogen solution was tested. The accuracy of each model is obtained from the spectral scan, and after comparison, the advantages offered by RF modeling are judged and analyzed.

PLS, SVM and RF were used to model the concentration-absorbance data of the bands corresponding to the two parameters, and the correlation coefficient R^2 and the *RMSE* of the predicted values of each model were compared. The specific equations are as follows

$$\bar{x}=\frac{1}{n}\sum\nolimits_{i=1}^{n}x(i);\quad \bar{x}'=\frac{1}{n}\sum\nolimits_{i=1}^{n}x'(i), \tag{1}$$

$$R=\frac{\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)\left(x'(i)-\bar{x}'\right)}{\left[\sqrt{\sum_{i=1}^{n}\left(x(i)-\bar{x}\right)^2}\right]\times\left[\sqrt{\sum_{i=1}^{n}\left(x'(i)-\bar{x}'\right)^2}\right]}, \tag{2}$$

$$RMSE=\sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}[x(i)-x'(i)]^2}, \tag{3}$$

where x is the raw data, x' is the data after modeling, and n is the number of data.

NO_3-N has a peak between 210—220 nm, which is modeled by PLS. Construct 6 groups of NO_3-N standard solutions with different concentrations of 0—10 mg·L^{-1}, take the absorbance value every 5 nm in 220—240 nm as the dependent variable, and the concentration value of the standard solution as the independent variable to establish

NO_3-N-PLS algorithm model. The comparison between the predicted value and the actual value is shown in Fig.5.

Divide the 210—240 nm absorbance data of 20 groups of NO_3-N solutions into 15 groups as training sets and 5 groups as prediction sets, with penalty factor C=21.112 1, kernel function parameter g=0.027 204 7, and establish NO_3-N-SVM algorithm model. The comparison between the predicted value and the actual value is shown in Fig.6.

Divide the 210—240 nm absorbance data of 20 groups of NO_3-N solutions into 15 groups as training sets and 5 groups of prediction sets, the number of decision trees is N_{tree}=500, the number of features is m_{try}=16, and the NO_3-N-RF algorithm model is established. The comparison between the predicted value and the actual value is shown in Fig.7.

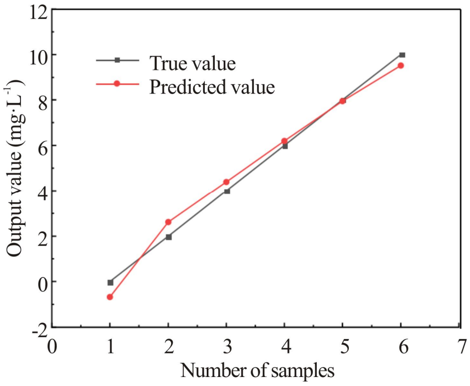

Fig.5 Comparison between the predicted value and the actual value for the NO_3-N-PLS model

The correlation coefficient of NO_3-N-PLS model is R^2=0.982 1, and $RMSE$ = 0.209 0. The fitted equation is shown as

$$y_{NO_3} = -0.669\,381 + 1.751\,352 x_1 + 2.279\,695 x_2 + 4.774\,685 x_3 + 10.750\,379 x_4 + 33.481\,249 x_5, \quad (4)$$

where x_1, x_2, x_3, x_4 and x_5 are the absorbance at 220 nm, 225 nm, 230 nm, 235 nm and 240 nm, respectively. In the NO_3-N-SVM model, the correlation coefficient is R^2=0.980 9, and $RMSE$=0.699 7. In the NO_3-N-RF model, the correlation coefficient is R^2=0.995 1, and $RMSE$=0.451 6.

NO_2-N has a peak near 210 nm, which is modeled by PLS. Construct 6 groups of NO_2-N standard solutions with different concentrations of 0—10 mg·L^{-1}, take the absorbance value every 5 nm in 205—225 nm as the dependent variable, and the concentration value of the standard solution as the independent variable to establish NO_2-N-PLS model. The comparison between the predicted value and the actual value is shown in Fig.8.

Divide the 210—240 nm absorbance data of 20 groups of NO_2-N solutions into 15 groups as training sets and 5 groups as prediction sets, with penalty factor C=337.794, kernel function parameter g=0.008 974 21, and establish the NO_2-N-SVM algorithm model. The comparison between the predicted value and the actual value is shown in Fig.9.

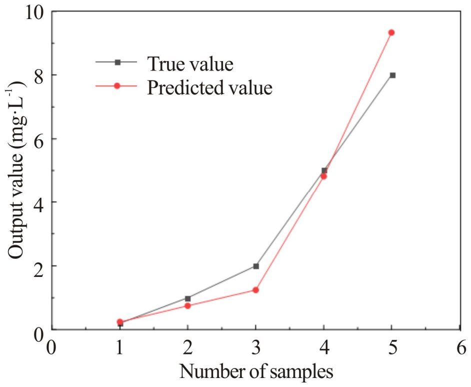

Fig.6 Comparison between the predicted value and the actual value for the NO_3-N-SVM model

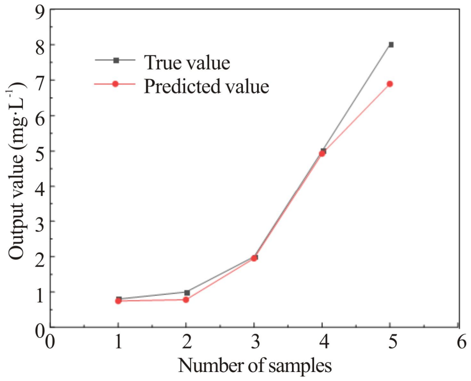

Fig.7 Comparison between the predicted value and the actual value for the NO_3-N-RF model

Divide 20 groups of NO_2-N solution's 205—230 nm absorbance data into 15 groups as training sets and 5 groups as prediction sets, the number of decision trees is N_{tree}=500, the number of features is m_{try}=13, and the NO_2-N-RF algorithm model is established. The comparison between the predicted value and the actual value is shown in Fig.10.

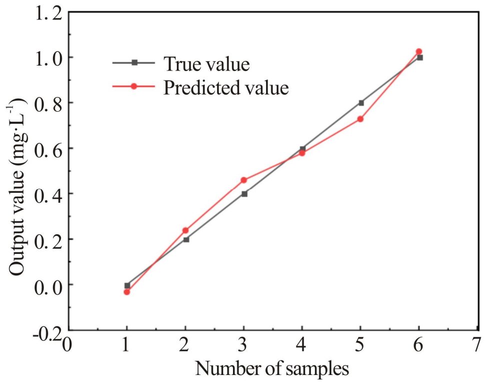

Fig.8 Comparison between the predicted value and the actual value for the NO_2-N-PLS model

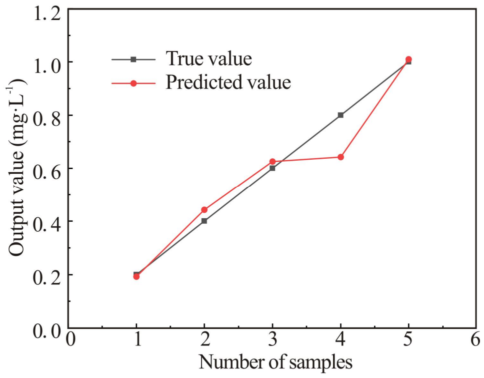

Fig.9 Comparison between the predicted value and the actual value for the NO_2-N-SVM model

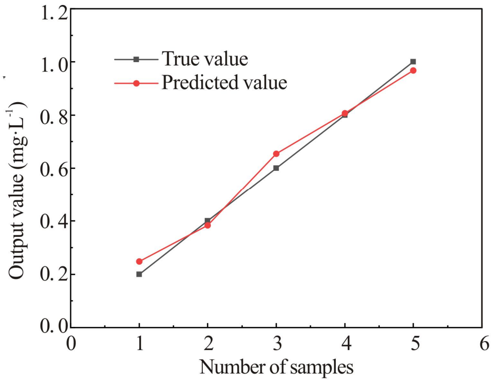

Fig.10 Comparison between the predicted value and the actual value for the NO_2-N-RF model

The correlation coefficient of the NO_2-N-PLS model is R^2=0.982 7, and *RMSE*=0.450 0. The fitted equation is shown as

$$y_{NO_2} = -0.027\,795 + 0.636\,735x_1 + 0.650\,653x_2 + 0.615\,890x_3 + 0.774\,200x_4 + 0.969\,458x_5, \quad (5)$$

where x_1, x_2, x_3, x_4 and x_5 are the absorbance at 205 nm, 210 nm, 215 nm, 220 nm and 225 nm, respectively. In the NO_2-N-SVM model, the correlation coefficient is R^2=0.935 3, and *RMSE*=0.074 2. In the NO_2-N-RF model, the correlation coefficient is R^2=0.994 2, and *RMSE*=0.036 5.

The prediction evaluation model is evaluated based on the correlation coefficient R^2 and *RMSE*. The more R^2 converges to 1 and the smaller the *RMSE*, the better the predicted data overlap with the original data and the better the regression model.

The modeling methods with PLS, SVM, and RF algorithms are presented above, and tables are constructed respectively for the models of both NO_3-N and NO_2-N parameters, as well as the correlation coefficient R^2 and *RMSE* values of the predicted and true values in the prediction set, as shown in Tabs.1 and 2.

Comparison of the results of the three methods in the table shows that the RF model has the highest correlation coefficient R^2 and the best effect, and the *RMSE* is lower compared with the other two methods, so the RF algorithm applied to water quality detection NO_3-N and NO_2-N modeling prediction effect is better than the other two methods.

To ensure the accuracy of the conclusion, the experiment was repeated three times as above with NO_2-N as the object. It is proved that the *RMSE* values for both the NO_2-N-RF model and the NO_2-N-SVM model meet the test criteria, and the R^2 for the NO_2-N-RF model is greater than that for the NO_2-N-SVM model. The results are shown in Tab.3.

Tab.1 NO_3-N solution modeling results

NO_3-N model	Model parameters		Results	
			R^2	*RMSE*
PLS	Eq.(4)		0.982 1	0.209 0
SVM	C=21.112 1	g=0.027 2	0.980 9	0.699 7
RF	N_{tree}=500	m_{try}=16	0.995 1	0.451 6

Tab.2 NO_2-N solution modeling results

NO_2-N model	Model parameters		Results	
			R^2	*RMSE*
PLS	Eq.(5)		0.982 7	0.450 0
SVM	C=337.794 3	g=0.009 0	0.935 3	0.074 2
RF	N_{tree}=500	m_{try}=13	0.994 2	0.036 5

Tab.3 Comparison of NO_2-N experimental results

NO_2-N	R^2 (RF)	R^2 (SVM)	Difference	Lift ratio (%)	*RMSE* (RF)	*RMSE* (SVM)
Group 1	0.994 2	0.935 3	0.058 9	6.297 4	0.036 5	0.074 2
Group 2	0.997 3	0.987 7	0.009 6	0.971 9	0.034 8	0.010 7
Group 3	0.997 6	0.981 0	0.016 6	1.692 1	0.039 5	0.016 1
Group 4	0.996 8	0.970 3	0.026 5	2.731 1	0.037 1	0.027 2

In this paper, NO_3-N and NO_2-N standard solutions were used as the parameters to be measured, the regression of concentration and absorbance was established after noise reduction of the data scanned by the spectral detector using the RF algorithm, and the PLS and SVM algorithm models were compared to establish an inverse model with feasibility and reliability. After the inverse model comparison, the estimation of NO_3-N and NO_2-N concentrations by RF algorithm can improve the accuracy of the prediction model. The results of several experiments show that the RF algorithm predicts NO_2-N concentrations significantly better than the SVM algorithm and PLS algorithm. The inverse model established by continuous spectral water quality online detection technology under RF algorithm solves the problem of severe covariance between variables, and has good predictive ability for the concentration of the parameters to be measured in water samples, with high model accuracy, good adaptability and feasibility.

Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

References

[1] WEI K L, CHEN M, WEN Z Y. Research on signal processing for water quality monitoring based on continuous spectral analysis[J]. Spectroscopy and spectral analysis, 2014, 34(12)：3368-3373.

[2] WANG C L, WANG B, JI T. Simulated estimation of nitrite content in water based on transmission spectrum[J]. Spectroscopy and spectral analysis, 2022, 42(07)：2181-2186.

[3] HE M X, LI J, FAN W Y. Correlation between floc morphology and water quality based on partial least squares[J]. The administration and technique of environmental monitoring, 2021, 33(06)：48-51.

[4] YAN W L, REN S Y, YUE X X. Rapid detection of cAMP content in red jujube using near-infrared spectroscopy[J]. Optoelectronics letters, 2018, 14(5)：380-383.

[5] WANG Y M, CHEN H R, CHEN J Y. Comparation of rice yield estimation model combining spectral index screening method and statistical regression algorithm[J]. Transactions of the Chinese society of agricultural engineering, 2021, 37(21)：208-216.

[6] CASTRILLO M, GARCÍA L Á. Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods[J]. Water research, 2020, 172(C).

[7] MU H Y. Multi-models combined water quality analyzing based on multi-spectra[D]. Hangzhou：Zhejiang University, 2011.

[8] LI R N, WANG Q, LIU S M. Water quality warning method based on canonical correlation coefficient and random forest[J]. China environmental science, 2021, 41(09)：4457-4464.

[9] MAHSA M, KHANMOHAMMADI K M, HOSSEIN G. Classification of nanofluids solutions based on viscosity values：a comparative study of random forest, logistic model tree, Bayesian network, and support vector machine models[J]. Infrared physics and technology, 2022, 125：104273.

[10] NAFOUANTI B N, LI J X, ABBA M N. Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China：comparison of random forest, logistic regression and artificial neural network[J]. Applied geochemistry, 2021, 132：105054.

[11] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1)：5-32.

[12] YUAN Z X. Study on spectral classification model based on random forest[J]. Modern information technology, 2021, 5(07)：81-84.

[13] LI S F, JIA M Z, DONG D M. Fast measurement of sugar in fruits using near infrared spectroscopy combined with random forest algorithm[J]. Spectroscopy and spectral analysis, 2018, 38(06)：1766-1771.

[14] WANG K, WANG J X, XING Z N. Infrared spectrum modeling method based on RF algorithm of improved feature selection[J]. Application research of computers, 2018, 35(10)：3000-3002.

[15] VLI W, LV B B, FU H. Study on denoising of continuous spectrum on-line monitoring signal of water quality with micro-reagents based on HHT[J]. Optoelectronics letters, 2022, 18(2)：115-121.

[16] FANG T J, AMIE A, SHANEEL C. Electrochemical detection of nitrate, nitrite and ammonium for on-site water quality monitoring[J]. Current opinion in electrochemistry, 2022, 32：100926.

[17] MELISSA T, ALAN K. Assessing the accuracy of nitrate concentration data for water quality monitoring using visual and cell phone quantification methods[J]. Citizen science：theory and practice, 2021, 6(1)：2.

[18] DONG C Y, LI W Z, WANG Z H. An automated flow-batch analyzer based on spectrophotometry for the determination of nitrite[J]. Optoelectronics letters, 2019, 15(5)：339-342.