Gaussian Process Regression for Prediction of Sulfate Content in Lakes of China

Jingying Zhao[1,2], Hai Guo[2*], Min Han[1], Haoran Tang[2] & Xiaoniu Li[2]

[1]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116023, China
[2]College of Computer Science and Engineering, Dalian Minzu University, 18 LiaoheWest Road, Dalian Development Zone, Dalian 116600, China
*E-mail: guohai@dlnu.edu.cn

Abstract. In recent years, environmental pollution has become more and more serious, especially water pollution. In this study, the method of Gaussian process regression was used to build a prediction model for the sulphate content of lakes using several water quality variables as inputs. The sulphate content and other variable water quality data from 100 stations operated at lakes along the middle and lower reaches of the Yangtze River were used for developing the four models. The selected water quality data, consisting of water temperature, transparency, pH, dissolved oxygen conductivity, chlorophyll, total phosphorus, total nitrogen and ammonia nitrogen, were used as inputs for several different Gaussian process regression models. The experimental results showed that the Gaussian process regression model using an exponential kernel had the smallest prediction error. Its mean absolute error (MAE) of 5.0464 and root mean squared error (RMSE) of 7.269 were smaller than those of the other three Gaussian process regression models. By contrast, in the experiment, the model used in this study had a smaller error than linear regression, decision tree, support vector regression, Boosting trees, Bagging trees and other models, making it more suitable for prediction of the sulphate content in lakes. The method proposed in this paper can effectively predict the sulphate content in water, providing a new kind of auxiliary method for water detection.

Keywords: *environmental monitoring; Gaussian process regression; machine learning; sulphate content; water quality modeling.*

1 Introduction

As human activity changes the properties and tissue of natural water, it also affects the use value of water and endangers human health by water pollution [1]. Water pollution mainly refers to the phenomenon of pollutants discharged by human activities entering a water body and causing the water quality to decline and the use value to decrease or vanish. There are two categories of water pollution causes: the first are human factors, mainly comprising industrial waste water but also domestic sewage, drainage of farmland, pollutants in the atmosphere and rubbish deposited in the ground that are leached by rainfall and

Received September 20th, 2018, Revised January 15th, 2019 , Accepted for publication February 19th, 2019.
 ISSN: 2337-5779, DOI: 10.5614/j.eng.technol.sci.2019.51.2.4

end up in the water. The second category comprises natural factors, such as weathering and hydrolysis of rocks, volcanic eruptions, water erosion of the ground, precipitation leaching from atmospheric dustfall, etc. The substances released by organisms (mainly green plants) in the geochemical cycle are all sources of natural pollutants. Since human factors account for the majority of water pollution it is usually stated that water pollution is caused by human factors [2-3].

Sulfate is widely distributed in nature. The concentration of sulfate in natural water can range from several mg/L to several kg/L. Sulfate in surface water and groundwater mainly comes from weathering and leaching of mineral components into rock soil. Oxidation of metal sulfide also increases the sulfate content. Sulfate can damage the soil structure, reduce soil fertility, and adversely affect water systems. Sulfate content is an important parameter in water quality monitoring, especially in the monitoring of groundwater and tap water. Hence, it is very important to monitor and forecast sulfate content. China is a country with a large number of lakes, especially in the middle and lower reaches of the Yangtze River. Monitoring and forecasting of water quality in this area is essential. Water quality measurement and monitoring include physical and chemical detection methods and automatic sensor testing methods [4-5].

In recent years, the research on water quality monitoring and prediction has become a hot topic in academic circles. Naubi, *et al*. studied the water quality of the Skudai River and analyzed and determined the pollution level of the Skudai River based on spatial variation trends of the water quality index (WQI) and its sub-indexes. At the same time, the water quality of the Skudai River was evaluated by conductivity, turbidity, temperature, total dissolved solids, total phosphorus and nitrogen [6]. Cloete, *et al*. used smart sensors to design a real-time water quality monitoring system that can measure the physical and chemical parameters of water quality, such as flow rate, temperature, pH value, electrical conductivity and redox potential. Their experimental results showed that the system can read the physical and chemical parameters in water and can process, transmit and display the data successfully [7]. Rachel, *et al*. assessed the fecal contamination testing programs of 72 agencies in 10 countries to assess the status of regulated water quality monitoring in sub-Saharan Africa. The assessment showed that smaller water providers and rural public health offices require greater attention and additional resources to achieve regulatory compliance for water quality monitoring in sub-Saharan Africa [8]. Shively, *et al*. proposed a beach water quality prediction model that sends water buoys and weather stations over wireless networks to servers, predicts them through empirical models and transmits the predictions to lifeguards at the beaches. Their experimental results showed that the prediction performance of this model

is better than that of persistence models and can effectively monitor beach water quality [9]. Kumpel, *et al*. proposed a water quality monitoring model that uses monitoring data to assess drinking water quality and water safety management in sub-Saharan areas. The experimental results showed that the level of fecal indicative bacteria (FIB) supplied by pipes was lower than that of any other source type. Real-time collection of water quality is very important for the safety of drinking water [10]. Partyka, e*t al*. sampled water quality in the Sacramento-San Joaquin Delta Estuary (Delta), created a baseline of microbial water quality in the Delta and identified various factors (climatic, land use, tidal, etc.), and used model prediction to analyze it. The experimental results showed that spatial auto-correlation was a major component of the water quality outcomes [11]. Wang, *et al*. developed a multi-sensor wireless intelligent water quality monitoring system that uses an STC12C5A60S2 micro-controller as the main control chip. It can remotely monitor and control pH, temperature, turbidity and other parameters in water. The experimental results showed that the system has high accuracy and can effectively reduce the consumption of manpower and financial resources [12].

In recent years, machine learning has been widely applied in various fields of environmental engineering [13-15]. Yang, *et al*. used a combination of dynamic principal component analysis and support vector machine to identify fault types and conflicts in a water quality monitoring and control (WQMC) system. The experimental results showed that the recognition accuracy of this method was 90%~94%. It could identify fault types and conflicts accurately and can be helpful in the maintenance and management of WQMC equipment [16]. Luna, *et al*. developed a water quality monitoring and water supply automation system for aquaculture. The experimental results showed that the system could effectively monitor the water quality and feed crayfish [17]. Ahmad, *et al*. proposed a multi-neural network model for real-time prediction of the BOD and COD water quality indexes and built a forecast sample of Perak River in Malaysia. The experimental results showed that the single feed-forward neural network model could predict WQI well, with coefficient of determination R2 and mean squared error (MSE) at 0.9090 and 0.1740 respectively.

Through multi model aggregation, the prediction error value MSE is lower than that of a single model that can effectively predict water quality [18]. Gebler, *et al*. proposed a river ecological state prediction model based on an artificial neural network. The model used physical and chemical parameters reflecting water quality and hydrological morphological characteristics as the explanatory variables of the artificial neural network and normalized root mean square error and coefficients of determination as evaluation indexes. The experimental results showed that the model could effectively reflect the water quality and hydrological morphology condition of rivers [19]. Khataar, *et al*. proposed a

method for predicting water quality using an artificial neural network, thus affecting the saturated hydraulic conductivity of soil. The neural network is trained by the Levenberg-Marquardt (LM) and Bayesian regulation algorithms. The salinity and alkalinity of the water are the inputs of the model, and saturated (Ks) and relative (Kr) hydraulic conductivities are the outputs. The experimental results showed that the method is superior to other linear regression methods [20]. Zhang, *et al*. proposed a short-term water quality prediction method based on multiple machine learning methods. In their method, dissolved oxygen, chemical oxygen demand by KMnO4 and ammonia nitrogen are used as the inputs of a support vector machine, and the optimal wavelet neural network based on particle swarm optimization algorithm is used to predict the overall state index of the water quality. The experimental results showed that the model is superior to the traditional BP neural network model, wavelet neural network model and gradient enhancement decision tree model [21].

Li, *et al*. proposed a method of predicting chlorophyll A in lake water with different water quality using hybrid neural networks. After clustering the water quality data of different lakes, the genetic algorithm optimized back-propagation neural network is used to predict the water quality. The experimental results showed that its prediction performance is good [22]. Liu, *et al*. proposed a fault diagnosis model based on multiclass support vector machines and rule-based decision trees for a water-quality monitoring device. The experimental results showed that the RBDT-MSVM algorithm could be effectively applied to fault diagnosis of a water quality monitoring device in a river crab breeding pond; the classification accuracy reached 92.86%, which is superior to other algorithms [23]. Chen, *et al*. proposed a new machine learning method, Support Function Machine, which was used in water quality evaluation. The experimental results showed that the method could effectively classify and evaluate water quality data [24]. Heddam, *et al*. used Least Square Support Vector Machine, Multivariate Adaptive Regression Splines, M5 model Tree and other machine learning methods to predict the dissolved oxygen concentration in water. This method takes water temperature, pH, specific conductance and discharge as input data and inputs them into three respective models. The experimental results showed that the three models had the best prediction performance for dissolved oxygen in water and the prediction accuracy of the three models was different at different stations [25]. Wu, *et al*. used a modular artificial neural network (MANN) and data preprocessing by singular spectrum analysis (SSA) to eliminate the lag effect. The experimental results showed that SSA could considerably improve the performance of the prediction model and eliminate the lag effect, and the ANN R-R model coupled with SSA was the most promising [26]. Cheng, *et al*. have proposed a parallel genetic algorithm with a fuzzy optimal mode that can significantly reduce the

overall optimization time and simultaneously improve the solution quality [27]. Taormina has proposed a binary-coded swarm optimization and Extreme Learning Machines. The results showed that there is no evidence that MM outperforms global GM for predicting total flow [28].

In the present study, water temperature, transparency, pH, dissolved oxygen, conductivity, chlorophyll, total phosphorus, total nitrogen and amino nitrogen were used as inputs to predict the sulphate content in water by Gaussian process regression. This model can replace the traditional measurement method of sulphate content in water, which is a manual physical and chemical method. It can effectively reduce the consumption of manpower and financial resources. In the following, data and variable selection, Gaussian process regression modeling and model performance assessment are introduced first. The model selection and a method comparison are described next.

Five-fold cross validation is also presented for discussing the accuracy of the Gaussian process regression with exponential kernel function model and other Gaussian kernel function prediction models. At the same time, it was compared with Linear-SVR (Support Vector Regression), Quadratic-SVR (Support Vector Regression), RBF-SVR, (Support Vector Regression) Boosting trees and Bagging trees. Finally, the paper concludes with a summary of this paper.

2 Methodology

2.1 Data and Variable Selection

In order to verify the prediction model of carbon content in water, this study used Python to write the crawlers. Other programming languages could have been used for programming the crawlers, such as C, C++, JAVA, C#, etc. but their capture data results are the same. In the early development stages of crawlers, C, C++ and JAVA were widely used, but in recent years, almost all crawlers are written in Python, because Python has a large number of built-in class libraries, making the program easier to write. Therefore, the crawlers in this study were written in the Python language.

The water quality monitoring data of the middle and lower reaches of the Yangtze River from 2007 to 2009 were obtained from the Lake-basin Thematic Library for the Middle and Lower Reaches of the Yangtze river (http://www.lakesci.csdb.cn/front/detail-lake2014zdhpszrgjc?id=2000) in the Chinese Lake Database. 515 water-quality monitoring samples were selected as experimental samples. Water temperature, transparency, pH, dissolved oxygen, conductivity, chlorophyll, total phosphorus, total nitrogen and ammonia nitrogen as the inputs of the model, are denoted as X1, X2, X3, X4, X5, X6, X7,

X8, X9, and the sulphate content in the water as the output of the model, is denoted as Y. The details are shown in Figure 1.

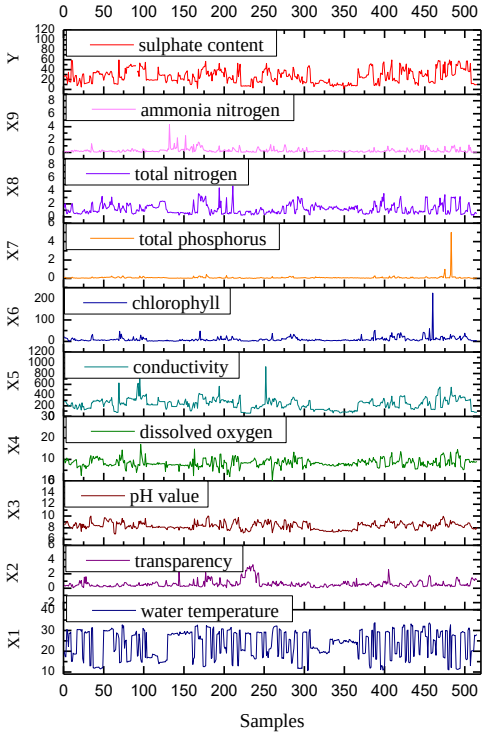

Figure 1 Prediction sample library.

2.2 Gaussian Process Regression Modelling

Gaussian process regression is a machine learning method based on statistical learning theory and Bayesian theory. It is suitable for dealing with complex regression problems, such as high dimensions, small sample sizes and non-linearity, and it has strong generalization ability. Compared with neural networks and support vector machines, Gaussian process regression has many advantages, such as easy realization, self-adaptive acquisition of hyper-parameters, flexible inference of non-parameters and the probabilistic significance of its output. In statistics and machine learning, some basic theories and algorithms are universal, but the basic concern of statistics is to understand the relationship between data and a model, and the main goal of machine learning is to predict more accurately and to better understand the behavior of the learning algorithms. Machine learning is a black box algorithm and using statistics is more likely to get the theoretical interpretation of the model. A Gaussian process model links statistics with machine learning at some level.

Gaussian processes are mathematically equivalent to many well-known models, including the Bayes linear model, Spline model, neural networks under suitable conditions, and Gaussian processes are also closely related to support vector machines [29-31].

Random processes can be represented by a cluster of random variables. What distinguishes Gaussian processes from other random processes is that the joint distribution of the vectors of the variables obtained by arbitrarily extracting a finite number of indicators in this random variable cluster is a multidimensional Gaussian distribution. In a Gaussian process, each point in the input space is associated with a random variable that obeys the Gaussian distribution and the joint probability of any finite number of these random variables also obeys the Gaussian distribution. When the indicator vector t is two-dimensional or multidimensional, the Gaussian process becomes a Gaussian random field. The characterization of a Gaussian process is like the characterization of a Gaussian distribution, which is also characterized by means and variance. In the application of Gaussian processes, the mean m is assumed to be zero, while the covariance function K is determined according to the specific application. In the Gaussian process, the parametric model is discarded and the prior probability distribution on the function is defined directly. In Gaussian process regression, it is not necessary to specify the specific form of the function, the observed values of N training data are considered to be a point(n-dimension) sampled from a multidimensional(n-dimension) Gaussian distribution without specifying the specific form of the function. Similarly, it can also be considered to be an infinite-dimension point sampled from a Gaussian process [32].

As early as 1964, Aizermann, *et al*. introduced this technology in the field of machine learning in a study of the potential function method, but its potential was not fully exploited until 1992 when Vapnik, *et al*. successfully extended linear SVMs to nonlinear SVMs using this technology. The theory of the kernel function is much older. The Mercer theorem can be traced back to 1909, while the study of Reproducing Kernel Hilbert Space began in the 1940s. In general, kernel functions applicable to SVMs can also be applied to Gaussian process regression [33-35]. The kernel functions commonly used in Gaussian processes include Constant kernel, Exponential kernel, Matern 5/2 kernel, Squared Exponential kernel and Rational Quadratic kernel, as shown in Eqs. (1) to (5):

Rational quadratic kernel

$$k(x,y) = (1 + \frac{d(x,y)^2}{2\alpha l^2})^{-\alpha} \tag{1}$$

Squared exponential kernel

$$k(x,y)=\exp(-\frac{\|x-y\|^2}{2\sigma^2}) \tag{2}$$

Constant kernel

$$k(x,y)=constant\_value\forall x,y \tag{3}$$

Exponential kernel

$$k(x,y)=\exp(-\frac{\|x-y\|}{2\sigma^2}) \tag{4}$$

Matern 5/2 kernel

$$k(x,y)=\sigma^2(1+\gamma\sqrt{5}d(\frac{x}{l},\frac{y}{l})+\frac{5}{3}\gamma^2 d(\frac{x}{l},\frac{y}{l})^2)\exp(-\gamma\sqrt{5}d(\frac{x}{l},\frac{y}{l}))\quad v=\frac{5}{2} \tag{5}$$

According to the different kernel functions, Gaussian process regression models with different kernel functions were designed to predict the sulfate content in water. The prediction model is shown in Figure 2.

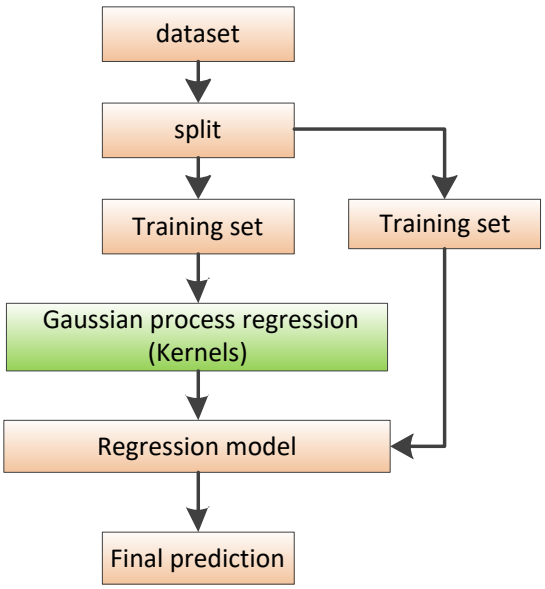

Figure 2 Gaussian regression models.

2.3 Model Performance Assessment

Many assessments can be used to assess the prediction performance, such as MAPE (mean absolute percentage error), MAE (mean absolute error), MSE

(mean squared error), RMSE (root mean squared error), determination coefficients R^2, R_2^2, etc. MAPE, MAE, MSE and RMSE can all be used to measure the error statistics of the prediction model and calculate the difference between the predicted value and the actual result of the regression model. The smaller the difference, the better the performance. MAE and RMSE were selected here. You can use all four but that would be redundant, because when the MAE and RMSE of a model are small, the MAPE and MSE are correspondingly small. When a model is suitable for a sample, its determination coefficients R2 and R_2^2 are between 0 and 1. The closer to 1, the higher the model's accuracy. The main consideration of these two quantities is to prevent errors in individual samples in the model from being too large to keep the average error low. In this study, two angles were selected to analyze the error: one is the mean error measurement using MAE and RMSE, and the other is determination coefficient R^2 to prevent individual sample error or the error ratio becoming too large. In this study, we used mean absolute error (MAE), root mean squared error (RMSE) and R-square (R^2) evaluation indexes.

The mean absolute error (MAE) is the difference between the predicted value and the measured value, which is inversely proportional to the prediction accuracy. The expressions are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right| = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right| \tag{6}$$

f_i is the predicted value and y_i is the measured value.

The RMSE value is inversely proportional to the prediction effect. The smaller the value, the higher the accuracy of the predictor.

$$RMSE = \sqrt{\frac{\sum_{n-1}^{n}\left(f_i - y_i\right)^2}{n-1}} \tag{7}$$

The value of R^2 is generally between 0 and 1. The closer R^2 is to 1, the smaller the prediction error of the model, the more accurate the prediction is. The expression of R^2 is as follows:

$$R^2\left(y, \hat{y}\right) = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{8}$$

Among them, the prediction response mean is $\bar{y} = \sum_{i=1}^{n} y_i$.

3 Results and Discussion

3.1 Model Selection

The specifications of the computer used in this research are Intel i5 8400 CPU and 16GB DDR-RAM. The programming software used in this research was Matlab. Several programming languages can be used to write regression models, such as C, C++, JAVA, C#, etc. Matlab was used to write the regression model because it makes it easier to implement the model. Five-fold cross validation was used for accuracy testing. Cross-validation is a widely used method for evaluating regression models and classification models.

In addition to 5-fold cross validation, 2-fold cross validation or k-fold cross validation can be used. This depends mainly on the amount of data processed by the computer and the complexity of the model. For large samples, the 2-fold cross validation method is used. Cross validation, sometimes also called rotation estimation, is a practical way to statistically cut data samples into smaller subsets, proposed by Seymour Geisser. The k-fold cross validation was used to randomly divide the sample set into *k* parts; *k*-1 parts of this sample set were used as training sample and 1 part was used as verification sample, and then the training and verification samples were rotated *k* times. This effectively reduces the overfitting risk of the prediction model.

The advantage of this method is that it repeatedly uses randomly generated sub-samples for training and verification. Each time the results are verified, the purpose of cross-validation is to obtain a reliable and stable model, to prevent artificially dividing test sets and training sets to be more precise. The effectiveness of the evaluation model. The parameters of this model are mainly kernel function selection, Sigma, Beta and Alpha. In the proposed method, a constant kernel is used as the basic kernel of the Gaussian process regression, and new kernel functions are constructed by combining the exponential, Matern 5/2, squared exponential and rational quadratic kernel with the basic kernel, respectively.

The optimal parameter in the model is the approximate optimal solution of the model parameter, which is searched by random search according to RMSE. The prediction error of the Gaussian process regression was the smallest when the exponential kernel function was used. Its main parameters were Sigma L = 7.99 2452542975736, Sigma F = 21.178416855950580, Beta = 30.4815941293384 15. Alpha was selected in the list. Figure 3 shows a comparison between the predicted value of the Gaussian process regression and the actual measured values. Table 1 shows an accuracy comparison of Gaussian processes using four different kernel functions.

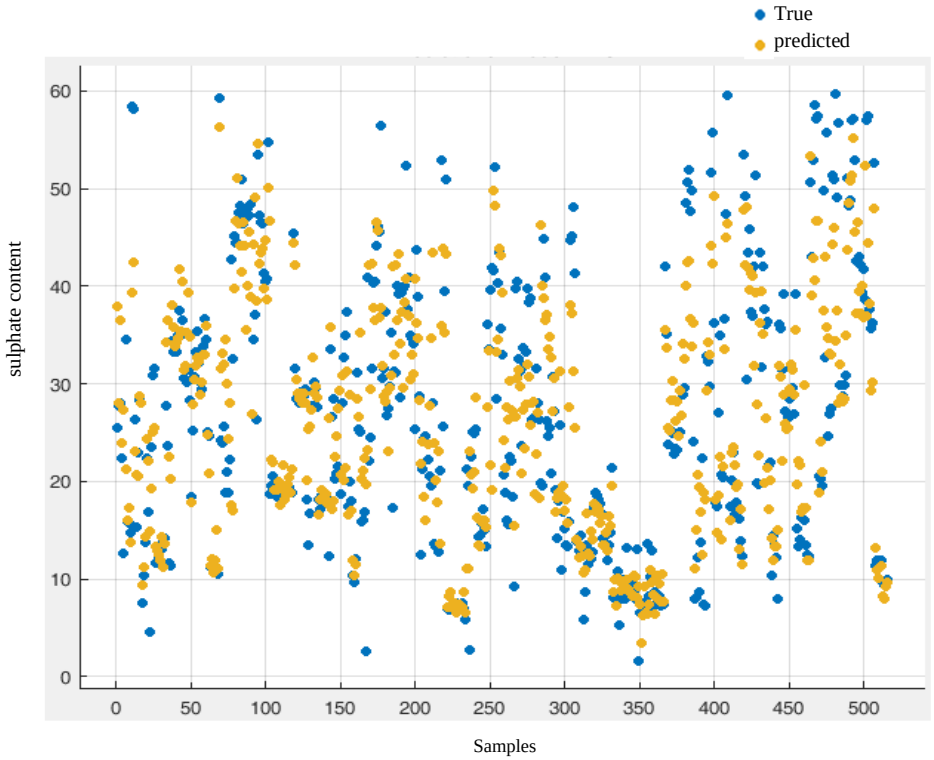

Figure 3 Comparison between predicted values and measured values.

Table 1 Comparison of prediction errors using different kernel functions.

kernel function	MAE	RMSE	R^2
Exponential	**5.0464**	**7.269**	**0.72**
Matern 5/2	5.4466	7.7305	0.69
Squared exponential	5.8015	8.1028	0.65
Rational quadratic	5.1177	7.3455	0.72

Through the error analysis shown in Figure 4, it was found that the MAE of the model was 5.0464 and the RMSE was 7.269 when the exponential kernel function was used, which is lower than that of Matern 5/2, squared exponential and rational quadratic. The R^2 of the model was 0.72 when the exponential kernel function was used, which is closer to 1. In conclusion, the Gaussian process regression with exponential kernel function was the most suitable for prediction of the sulphate content in lakes.

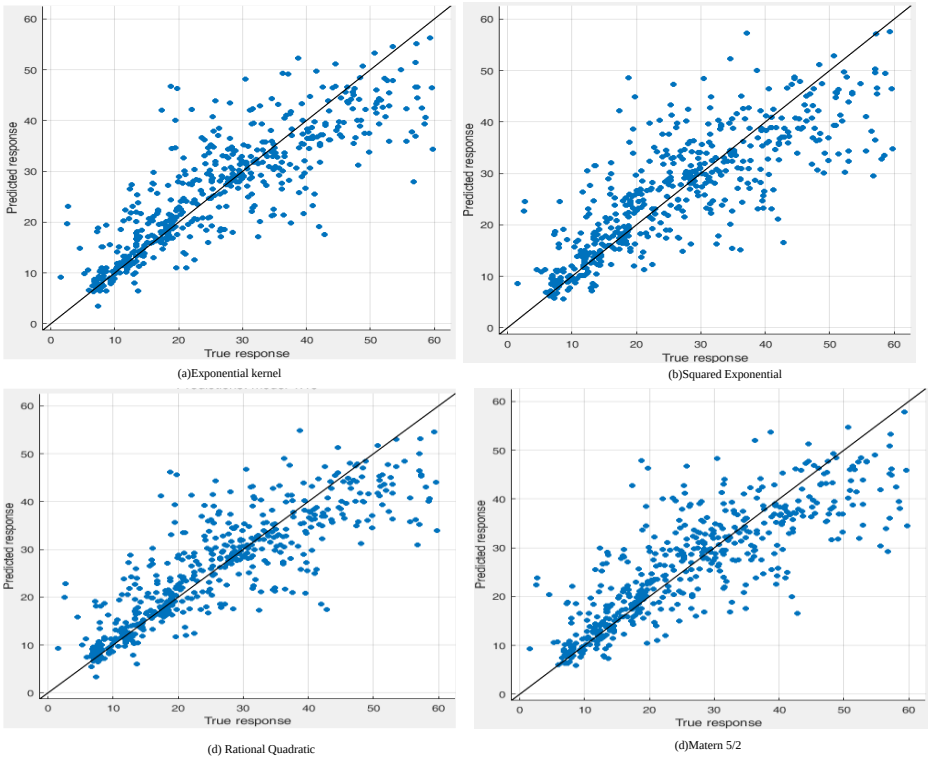

Figure 4 Prediction accuracy comparison of different kernel functions.

3.2 Method Comparison

In order to test the validity of the model, the model was compared with linear regression, decision tree, Linear-SVR (Support Vector Regression), Quadratic-SVR, RBF-SVR, Boosting trees and Bagging trees under the same experimental conditions. These are four main types of regression methods for machine learning. The first category is linear regression, which typically includes linear regression, ridge regression, lasso regression, etc., and its regression error is large, but the model training speed is fast. The second category is support vector regression, especially support vector regression with multiple kernel functions. The disadvantage here is that hyper-parameter acquisition is difficult and the training time is too long. The third category is the tree class, including the decision tree and its variants C3.1, C4.1, J48, etc. Their disadvantage is that the generalization ability is poor. The fourth category consists of integrated learning classes, such as bagging, boosting, and their variants. Their disadvantage is that they tend to cause overfitting when the sample is insufficient. Different regression models perform differently under different kinds of samples. In this paper, the typical linear regression, decision tree, Linear-SVR, Quadratic-SVR,

RBF-SVR, Boosting trees and Bagging trees were selected for the comparative experiments. The experimental results are shown in Table 2. Through the analysis of the experimental results, the prediction error of the model in this paper for sulphate content in water was MAE = 5.0464 and RMSE = 7.269, which was less than for linear regression, decision tree, Linear-SVR, Quadratic-SVR, RBF-SVR, Boosting trees and Bagging trees. The R^2 value of the prediction model was 0.72, which is closer to 1, and the prediction error of the model was smaller than that of the other models.

The analysis of Table 2 shows that the coincidence degree of the prediction points in the proposed model was better than that of linear regression, decision tree, Linear-SVR, Quadratic-SVR, RBF-SVR, Boosting trees and Bagging trees. Hence, this model is more suitable for predicting sulphate content in water than the other models.

Table 2 Error comparison of different models.

Regression model	MAE	RMSE	R^2
Exponential-GPR	**5.0464**	**7.269**	**0.72**
Linear regression	6.2266	8.8725	0.59
decision tree	6.1278	8.5228	0.62
Linear-SVR	6.0882	9.1003	0.56
Quadratic-SVR	6.355	14.106	-0.05
RBF-SVR	5.9864	8.2423	0.64
Boosting trees	5.4112	7.5405	0.70
Bagging trees	5.7213	7.7151	0.71

In the past, the sulphate content in water was measured by physical and chemical detection methods, and the efficiency was relatively low. In this paper, water temperature, transparency, pH, dissolved oxygen, conductivity, chlorophyll, total phosphorus, total nitrogen, ammonia nitrogen are used as input data. The data obtained by the sensor are fully utilized, so that the sulphate content in water can be predicted in real time, the cost of manual testing is reduced, and the detection efficiency is improved.

The method of predicting sulphate content in water based on Gaussian process regression is not only convenient and accurate, but more importantly, it can resolve the nonlinear relationship in complex systems more accurately, so that it can be realized under complex nonlinear conditions. More accurate real-time prediction of carbon content in water can effectively save test cost and time.

Under the same experimental conditions, using the 5 fold cross-validation method, the model's prediction error was smaller than that of linear regression, decision tree, linear SVR, Quadratic SVR, RBF-SVR, Boosting trees and Bagging trees. It was proved that under the same experimental conditions, the

accuracy based on Gaussian process regression forecasting model is higher than based on the other models and is more suitable for the sulphate content prediction in water. This study was only an attempt to predict the sulfate content in water from the perspective of technical methods. Taking the lakes in the middle and lower reaches of the Yangtze river as an example, the universality of the research results remains to be further discussed. It is believed that with the development of research and application, Gaussian process regression as an advanced artificial intelligence algorithm and a new prediction method can be more widely used in the prediction of other substances in water.

The application of Gaussian process regression model to accurately predict the sulphate content in water in a wide range of research areas is necessary for real-time analysis of water quality, especially for water quality monitoring in the middle and lower reaches of the Yangtze River. It is of great significance to promptly and accurately put forward effective measures to control water pollution, protect and restore water quality systems, improve water quality and realize sustainable utilization of water resources.

4 Conclusion

It is feasible to predict the sulphate content in water by using temperature, transparency, pH, dissolved oxygen, conductivity, chlorophyll, total phosphorus, total nitrogen, ammonia nitrogen as inputs, using sulphate content in the water as output, and using Matlab to compile a Gaussian process regression prediction model. Its prediction accuracy is high. The model was validated using a 5-fold cross validation method. The validation results showed that the prediction error of the carbon content in water was small, MAE was 5.0464, RMSE was 7.269, and R^2 was 0.72, which means the method can effectively predict the sulphate content in water.

Compared with linear regression, decision tree, Linear-SVR, Quadratic-SVR, RBF-SVR, Boosting trees, Bagging trees and other regression models, the prediction error of the model is smaller than that of these models, which makes it the most suitable for the prediction of the sulphate content in water.

The prediction model proposed in this paper can effectively use the water quality data obtained by a sensor network to predict the sulphate content in water in real time, reduce the cost of testing, and improve the detection efficiency. It can realize more accurate real-time predictions of the carbon content in water under complex non-linear conditions, and effectively save test costs and time. The next step is to improve the precision of the regression for three aspects: firstly, to build a larger sample and add data of other lakes to the sample in this paper; secondly, to design and verify the kernel function that is

most suitable for the sample in this paper; thirdly, to try to apply more machine learning methods to water quality prediction.

Acknowledgements

This article has obtained support from the National Natural Science Foundation of China (201501030401), the Foundation of Chinese Ministry of Education (18YJCZH040), the National Natural Science Foundation of Liaoning (201601084, 20170050, and 2018401030), the Fundamental Research Funds for the Central Universities, and the Doctoral Scientific Research Foundation of Dalian Minzu University. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers who improved the presentation.

References

[1] Han, D., Currell, M.J. & Cao, G., *Deep Challenges for China's War on Water Pollution*, Environmental Pollution, **218**, pp. 1222-1233, 2016.

[2] Xue, B.I., Yang, M. & Tian, Z., *Study on Water Pollution Characteristics of Huangbaihe River Watershed in Yichang and Its Control Measures*, Gastroenterology, **126**(3), pp. 290-300, 2017.

[3] Abdel-Satar, A.M., Ali, M.H. & Goher, M.E., *Indices of Water Quality and Metal Pollution of Nile River, Egypt*, Egyptian Journal of Aquatic Research, **43**(1), pp. 21-29, 2017.

[4] Spencer, M.A., Swallow, S.K. & Miller, C.J., *Valuing Water Quality Monitoring: A Contingent Valuation Experiment Involving Hypothetical and Real Payments*, Agricultural & Resource Economics Review, **27**(1), pp. 28-42, 2016.

[5] Behmel, S., Damour, M., Ludwig, R., Behmel, S. & Rodriguez, M.J. *Water Quality Monitoring Strategies - A Review and Future Perspectives*, Science of the Total Environment, **571**, pp. 1312-1329, 2016.

[6] Naubi, I., Zardari, N.H., Shirazi, S., Ibrahim, F. & Baloo, L., *Effectiveness of Water Quality Index for Monitoring Malaysian River Water Quality*, Polish Journal of Environmental Studies, **25**(1), pp. 231-239, 2016.

[7] Cloete, N.A., Malekian, R. & Nair, L., *Design of Smart Sensors for Real-Time Water Quality Monitoring*, IEEE Access, **4**, pp. 3975-3990, 2016.

[8] Rachel, P., Emily, K., Mateyo, B., Rahman, Z. & Khush, R., *To What Extent is Drinking Water Tested in Sub-Saharan Africa? A Comparative Analysis of Regulated Water Quality Monitoring*, International Journal of Environmental Research & Public Health, **13**(3), pp. 275, 2016.

[9] Shively, D.A., Nevers, M.B., Breitenbach, C., Phanikumar, M.S. & Przybyla-Kelly, K., *Prototypic Automated Continuous Recreational*

Water Quality Monitoring of Nine Chicago Beaches, Journal of Environmental Management, **166**, pp. 285-293, 2016.

[10] Kumpel, E., Peletz, R., Bonham, M. & Khush, R., *Assessing Drinking Water Quality and Water Safety Management in Sub-Saharan Africa Using Regulated Monitoring Data*, Environmental Science & Technology, **50**(20), pp. 10869-10876, 2016.

[11] Partyka, M.L., Bond, R.F., Chase, J.A. & Atwill, E.R., *Monitoring Bacterial Indicators of Water Quality in A Tidally Influenced Delta: A Sisyphean Pursuit*, Science of the Total Environment, **578**, pp. 346-356, 2017.

[12] Wang, W.H., Yue, W.G., Wang, Y.F., Wei-Hao, G. & Zhao, T.F., *Wireless and Intelligent Water Quality Monitoring System Design and Application with Multi-Sensor*, Electronic Design Engineering, **24**(7), pp. 135-140, 2016.

[13] Fotovatikhah, F., Herrera, M., Shamshirband, S., Chau, K.W., Ardabili, S. F. & Piran, M.J., *Survey of Computational Intelligence as Basis to Big Flood Management: Challenges, Research Directions and Future Work*, Engineering Applications of Computational Fluid Mechanics, **12**, pp. 411-437, 2018.

[14] Chau, K.W., *Use of Meta-Heuristic Techniques in Rainfall-Runoff Modelling*, Water, **9**(186), pp. 6, 2017.

[15] Wang, W.C., Xu, D.M., Chau, K.W. & Chen, S, *Improved Annual Rainfall-Runoff Forecasting Using PSO-SVM Model Based on EEMD*, Journal of Hydroinformatics, **15**, pp. 1377-1390, 2013.

[16] Yang, H., Hassan, S.G., Wang, L., Li, L. & Yang, H., *Fault Diagnosis Method for Water Quality Monitoring and Control Equipment in Aquaculture Based on Multiple SVM Combined with D-S Evidence Theory*, Computers & Electronics in Agriculture, **141**, pp. 96-108, 2017.

[17] Luna, F.D.V.B., Aguilar, E.D.L.R., Naranjo, J.S. & Jagüey, J.G., *Robotic System for Automation of Water Quality Monitoring and Feeding in Aquaculture Shadehouse*, IEEE Transactions on Systems Man & Cybernetics Systems, **47**(7), pp. 1575-1589, 2017.

[18] Ahmad, Z., Rahim, N.A., Bahadori, A. & Zhang, J., *Improving Water Quality Index Prediction in Perak River Basin Malaysia Through a Combination of Multiple Neural Networks*, International Journal of River Basin Management, **15**(1), pp. 79-87, 2017.

[19] Gebler, D., Wiegleb, G. & Szoszkiewicz, K., *Integrating River Hydromorphology and Water Quality into Ecological Status Modelling by Artificial Neural Networks*, Water Research, **139**, pp. 395-405, 2018.

[20] Khataar, M., Mosaddeghi, M.R., Chayjan R.A. & Mahboubi, A.A., *Prediction of Water Quality Effect on Saturated Hydraulic Conductivity of Soil by Artificial Neural Networks*, Paddy & Water Environment, **16**(3), pp. 631-641, 2018.

[21] Zhang, L., Zou, Z. & Shan, W., *Development of A Method for Comprehensive Water Quality Forecasting and Its Application in Miyun Reservoir of Beijing, China*, Journal of Environmental Sciences, **56**(6), pp. 240-246, 2017.
[22] Li, X., Sha, J. & Wang, Z.L., *Chlorophyll-A Prediction of Lakes with Different Water Quality Patterns in China Based on Hybrid Neural Networks*, Water, **9**(7), pp. 524, 2017.
[23] Liu, S., Xu, L., Li, Q., Zhao, X. & Li, D., *Fault Diagnosis of Water Quality Monitoring Devices Based On Multiclass Support Vector Machines and Rule-Based Decision Trees*, IEEE Access, **6**, pp. 22184-22195, 2018.
[24] Chen, J., Hu, Q., Xue, X., Ha, M. & Ma, L., *Support Function Machine for Set-based Classification with Application to Water Quality Evaluation*, Information Sciences an International Journal, **388**, pp. 48-61, 2017.
[25] Heddam, S. & Kisi, O., *Modelling Daily Dissolved Oxygen Concentration Using Least Square Support Vector Machine, Multivariate Adaptive Regression Splines and M5 Model Tree*, Journal of Hydrology, **559**, 2018.
[26] Wu, C.L. & Chau, K.W., *Rainfall-Runoff Modeling Using Artificial Neural Network Coupled with Singular Spectrum Analysis*, Journal of Hydrology, **399**, pp. 394-409, 2011.
[27] Cheng, C.T., Wu, X.Y. & Chau, K.W., *Multiple Criteria Rainfall-Runoff Model Calibration Using a Parallel Genetic Algorithm in A Cluster of Computer*, Hydrological Sciences Journal, **50**, pp. 1069-1087, 2005.
[28] Taormina, R., Chau, K.W. & Sivakumar, B., *Neural Network River Forecasting Through Baseflow Separation and Binary-Coded Swarm Optimization*, Journal of Hydrology, **529**, pp. 1788-1797, 2015.
[29] Datta, A., Banerjee, S., Finley, A.O. & Gelfand, A.E., *Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets*, Journal of the American Statistical Association, **111**(514), pp. 800-812, 2016.
[30] Kumar, S., Hegde, R.M. & Trigoni, N., *Gaussian Process Regression for Fingerprinting Based Localization*, Ad Hoc Networks, **51**, pp. 1-10, 2016.
[31] Gramacy, R.B. & Haaland, B., *Speeding Up Neighborhood Search in Local Gaussian Process Prediction*, Technometrics, **58**(3), pp. 294-303, 2016.
[32] Liu, J., Cuff, P. & Verdu, S., *Key Capacity for Product Sources with Application to Stationary Gaussian Processes*, IEEE Transactions on Information Theory, **62**(2), pp. 1146 - 1150, 2016.
[33] Guenther, N. & Schonlau, M., *Support Vector Machines*, Stata Journal, **16**(4), pp. 917-937, 2016.

[34] Andrew, A.M., *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Kybernetes, **32**(1), pp. 1-28, 2001.

[35] Moura, M.D.C., Zio, E., Lins, I. D. & Droguett, E., *Failure and Reliability Prediction by Support Vector Machines Regression of Time Series Data*, Reliability Engineering & System Safety, **96**(11), pp. 1527-1534, 2017.