Dean&Francis
ISSN 2959-6157

Development of a Lightweight Model for Short-term Prediction of COD Concentration under Limited Data Conditions

Yuxuan Ma

[1]Department of Enviroment Science, Northwest A&F University, Shaanxi, China
*Corresponding author: mayuxuan@arizona.edu

Abstract:

As a core indicator of the degree of organic pollution in water bodies, chemical oxygen demand (COD) monitoring results directly affect the identification of pollution events and the response efficiency of water quality management. Traditional detection methods are difficult to realize continuous high-frequency COD monitoring, resulting in limited timely response to pollution dynamics, which seriously impedes the initiative of water quality management. To address this problem, this paper explores the construction of a lightweight COD short-term prediction model with simple structure, low computational overhead and basic predictive ability based on real water quality monitoring data of the Weihe River Tieqiao section in Xianyang City, China, using three models: multiple linear regression, Lasso regression, and shallow decision tree. Through pre-processing and feature selection of water quality-related auxiliary variables, redundant information was eliminated, and finally pH, total nitrogen (TN), total phosphorus (TP), ammonia nitrogen (NH_3-N), and water temperature were identified as modeling variables. The results showed that the multiple linear regression model performed well in the medium concentration interval, but systematic bias existed in the high and low concentration intervals, reflecting the limitations of linear models for nonlinear features. The study in this paper provides a simple and effective modeling idea for the prediction of water quality indicators under resource-constrained conditions, which helps to improve the early warning capability and response efficiency in environmental management.

Keywords: Chemical Oxygen Demand (COD); Light-weight Prediction Model; Water Quality Monitoring.

1. Introduction

COD (Chemical Oxygen Demand) is a core indicator for measuring the degree of organic pollution in water bodies, and its monitoring results directly affect the identification of pollution events and the efficiency of water quality management response. Given its critical role in environmental management, timely and accurate monitoring of COD levels is essential for early warning systems and pollution control strategies. However, in practical environments, COD concentration often fluctuates dramatically. It changes rapidly, and the commonly used detection methods, whether standard chemical, spectroscopic, or online sensors, find it challenging to acquire continuous and high-frequency dat [1, 2]. Therefore, in daily monitoring, pollution changes often occur, but detection has not yet responded, and even the dynamics in the early stages of the trend cannot be captured. This limitation significantly hinders proactive water quality management. In this context, relying solely on traditional detection methods makes it difficult to achieve continuous observation and timely judgment of the COD evolution process [2]. Therefore, many studies have attempted to construct predictive models, using existing observational data to make short-term speculations on COD trends, to compensate for monitoring lag and enhance the foresight of water quality management.

A wide range of predictive models have been developed for COD forecasting, including artificial neural networks (ANN), partial least squares regression (PLS), and ensemble learning techniques. In multiple industrial wastewater cases, ANN has shown extremely high fitting accuracy (R^2 value can reach 0.9997), demonstrating strong nonlinear modeling capabilities [3]. However, the effectiveness of the prediction heavily depends on having sufficient and high-quality input data. When there are swift changes in concentration, the prediction errors tend to rise noticeably [4]. Ensemble learning methods—such as combinations of Random Forest and XGBoost—offer the advantage of integrating multiple variables and enhancing model robustness. However, their performance remains highly sensitive to the granularity and completeness of input datasets, thereby imposing greater demands on the precision and resolution of measurement parameters [5]. In contrast, traditional linear models such as MLR have meaningful disadvantages in accuracy (RMSE can reach 79.6 mg/L), but their structure is simple, computational cost is low, and they have deployment advantages [6]. These models have driven the advancement of COD prediction techniques from various angles, emphasizing the richness and technological progress possible through different algorithmic strategies in this area.

Although rich algorithmic achievements have been accumulated in the field of COD prediction, existing research still focuses on improving model performance and accuracy optimization, and there is a significant lack of exploration in model simplification and adaptability under resource constraints. Beyond neural networks, ensemble models such as random forests and XGBoost have gained attention. There is still a lack of systematic research on how to construct models with a simple structure, few parameters, and low computational overhead while ensuring basic predictive performance. Therefore, this article focuses on this relatively blank direction and explores whether a COD prediction model with a concise structure, acceptable prediction accuracy, and practical application potential can be constructed under limited data and resources.

2. Data collection and pre-processing

2.1 Data Source

This study selects the “Xianyang Iron Bridge” section of the Weihe River in Xianyang City as the research area, aiming to construct a short-term COD concentration prediction model under realistic monitoring conditions. The data used in this study is obtained from the official platform of the Ministry of Ecology and Environment of China—the Real-time Surface Water Quality Automatic Monitoring Data Publishing System. The dataset spans from January 2022 to May 2025, providing 6432 valid daily records and ensuring good continuity, consistency, and completeness of time series.

This study takes COD as the main predictive target variable. Select seven variables for data pre-processing based on the range of the original data: pH, dissolved oxygen (DO), ammonia nitrogen (NH_3-N), total nitrogen (TN), total phosphorus (TP), conductivity, and water temperature. These variables are consistently measured in water quality monitoring and theoretically linked to COD variation, either by affecting the degradation of organic matter or by indicating broader pollutant loading [7].

2.2 Pre-processing Method

Before establishing the model, this study systematically cleaned and preprocessed the raw data to ensure the quality and stability of the input variables. Firstly, convert the ‘Monitoring Time’ field to a timestamp format and set it as a table index to preserve its time series characteristics. The forward imputation method addresses missing values in the data, which replaces the current missing value with the valid value from the previous moment. The cleaning results showed 35 missing COD values, and the remaining

ISSN 2959-6157

variables were missing between 16-43. After all were successfully filled in, the data structure was complete, and the missing values were reset to zero.

Subsequently, the IQR (interquartile range) method was used to identify and eliminate outliers in the distribution, except for COD. One thousand six hundred seventy-seven samples were cleared, and 4755 valid data were retained. To make the variables comparable and reduce the impact of scale differences on the model results, all input variables except COD were Z-score standardized using StandardScaler in Python, with their mean close to 0 and standard deviation close to 1, to eliminate the influence of dimensional differences on the modeling effect. After completing standardization, merge the COD column into the dataset to obtain the final modeling data framework.

In addition, to provide a preliminary understanding of the distribution of auxiliary variables and subsequent feature selection, Matplotlib was used to generate correlation heatmaps and outlier boxplots for each variable. There was a significant collinearity between TN and EC (r=0.82), and water temperature also showed a strong negative correlation with TN and EC. To reduce the interference of redundant information and for the careful consideration of the explanatory power of pollution sources and data stability, TN and water temperature are retained in the model, while EC is excluded. The correlation between DO and COD is low, and the predictive value is limited. Many outliers in turbidity do not meet the purpose of lightweight in this article and need to be handled cautiously. The final selected modeling variables are pH, TN, TP, ammonia nitrogen, and water temperature, which will be used to construct the subsequent lightweight prediction model.

To improve data processing efficiency, some Python programming and preprocessing operations are completed under manual supervision using AI tools (ChatGPT), including the design of code logic suggestions, outlier recognition methods, and visualization methods. All steps are executed and verified by the author themselves to ensure the accuracy of code execution and result interpretation.

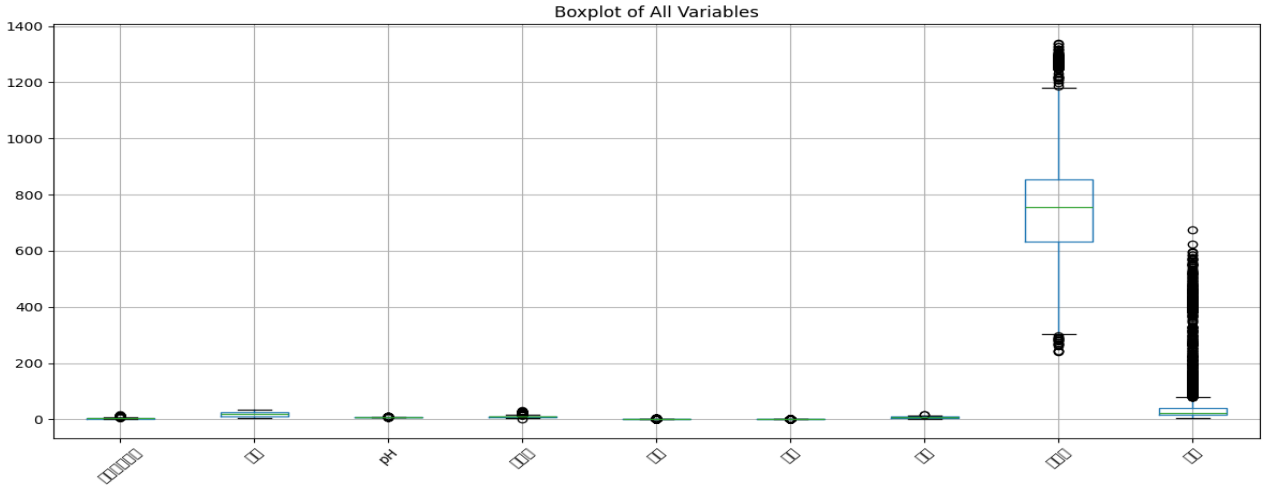

Fig. 1 Distribution of Outlier Box plots for Various Variables

Data from: National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform

Picture credit: Original

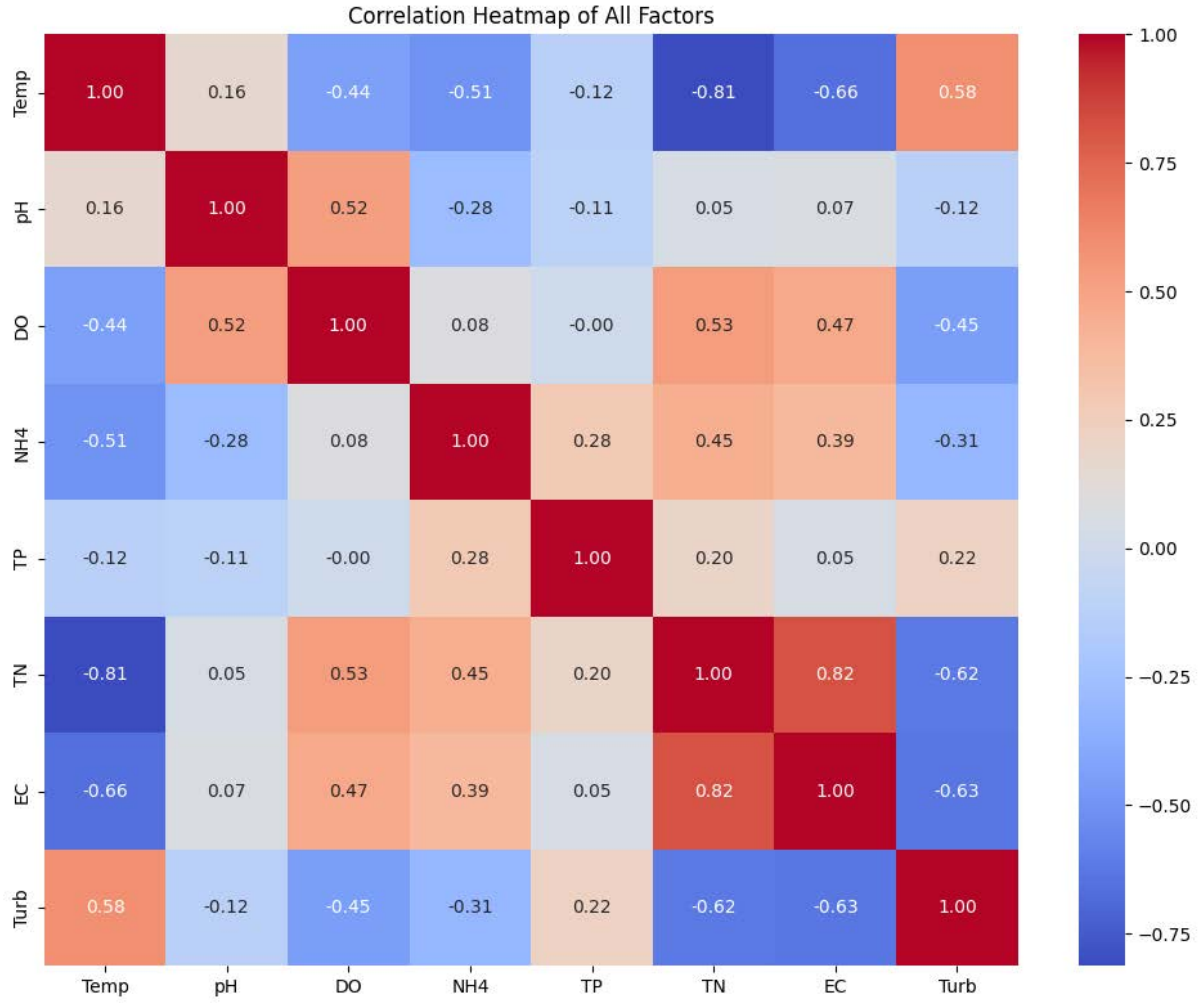

Fig. 2 Variable correlation heatmap

Data from: National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform

Picture credit: Original

3. Model building

3.1 Model Selection

To achieve the goal of "lightweight modeling," model selection should simultaneously meet three conditions: concise structure, low computational cost, and certain predictive performance, even under limited sample conditions. To satisfy the above constraints, this article selects three representative methods, namely multiple linear regression (MLR), Lasso regression, and shallow decision tree, for modeling and comparison.

As the most basic linear model, MLR has a clear structure and strong interpretability, making it suitable for establishing baseline performance references. Lasso regression introduces the L1 regularization term on this basis, which can automatically screen variables and remove redundant information while maintaining the model's explanatory power. The regression tree model that controls depth controls complexity while retaining a certain degree of nonlinear fitting ability, enhancing the model's flexibility. In addition, to further capture the trend of COD concentration changes in the time dimension, this paper introduces an AR model to model the sequence of historical data to explore the potential for short-term prediction based on temporal autocorrelation features. Each model is constructed based on the same training dataset and uniformly analyzed and evaluated using conventional error metrics.

3.2 MLR

3.2.1 Model Training

This model uses the least squares method to establish a linear relationship between COD and other water quality

ISSN 2959-6157

indicators. The input variables are five standardized variables: TN, TP, NH4, PH, and temperature, and the output variable is COD concentration. The data is divided into an 80% training set and a 20% testing set. The model is built on the Python platform and trained and predicted using the Linear Regression module in Sci-kit Learn. The parameters are kept at default settings, and no regularization or cross-validation is introduced.

3.2.2 Prediction Results

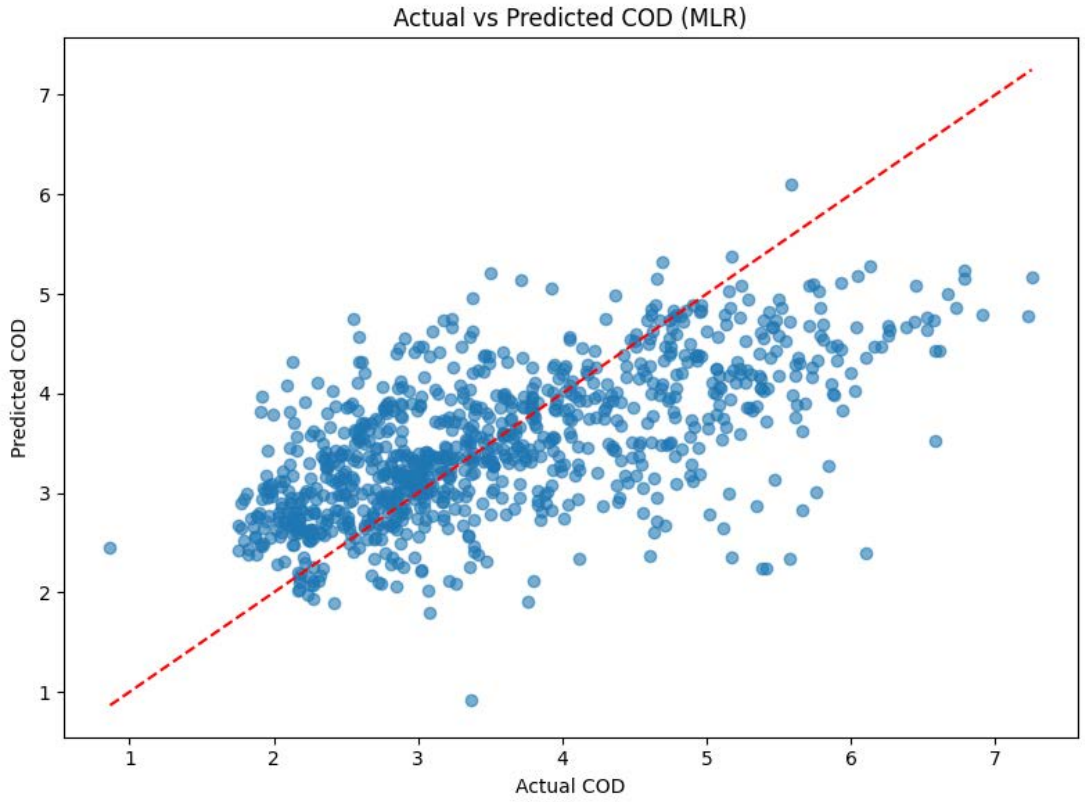

Fig. 3 MLR model prediction result chart

Data from: National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform

Picture credit: Original

Figure 3 shows the correspondence between the predicted results of the multiple linear regression model on the test set and the actual COD values. The scattered points are generally distributed around the ideal prediction line, but the overall point cloud structure shows the characteristics of "fitting in the middle and deviating at both ends." This indicates that the model's prediction effect is good in the medium concentration area, but the model's prediction error significantly increases in the high and low concentration areas.

In the mainstream range of COD concentration of 3-4 mg/L, the samples are dense, and the scatter points are mainly close to the ideal line. The model's predicted values are relatively close to the actual values, and the fluctuation amplitude is small. This result may be related to enough samples in the interval and the relatively uniform distribution of data, enabling the model to identify and fit patterns better. In the part where the COD concentration is higher than 5 mg/L, the scatter points shift downwards, generally below the ideal line, showing a systematic underestimation. This indicates that linear models cannot handle high-value intervals and may not be able to capture the nonlinear growth characteristics of pollutants at high concentrations. In addition, small low-concentration (<2 mg/L) samples showed mild overestimation, but the overall impact was relatively small.

The distribution trend of this prediction error may be influenced by two factors: first, the relatively small number of high-concentration samples leads to insufficient training of the model in this interval; second, the relationship between input variables and COD may have abrupt changes in highly polluted environments, and linear regression equations cannot explain this trend well.

3.2.3 Model Evaluation

The performance of the multiple linear regression model on the test set is generally weak, and the evaluation indicators are as follows:

R^2: 0.399, The model can only explain about 39.9% of the variation in COD concentration.
MAE: 0.692, The average deviation between each predicted value and the actual value is close to 0.7 mg/L.
RMSE: 0.903, Indicating significant fluctuations and extreme deviations in the error.
Combined with Fig.3, these indicators reflect that the model performs well in the middle concentration range but shows bias when dealing with extremely high or low COD levels. The overall fitting ability of the model is limited, and the error distribution structure has the typical characteristics of "concentrated in the middle and loose at both ends."
The fitting equation of this model is as follows:
$COD=3.447+0.379pH+0.747TN+0.258TP-0.061NH_4+0.501Temp$ (1)
From the regression coefficients, TN and water temperature have the highest weights in the model, at 0.747 and 0.501, respectively, indicating that they have the strongest linear explanatory power for COD. This result is consistent with the variable correlation in the previous heat map analysis, verifying the stable positive correlation between TN and COD. PH and TP also contribute positively, but the coefficient is relatively small. It is worth noting that the coefficient of NH_4 is negative and has the smallest absolute value, indicating its insufficient explanatory power in linear models. This may be due to its inconsistent direction of contribution to COD under different hydrological conditions or certain multicollinearity effects.
Although the model has a transparent structure, strong parameter interpretability, and is easy to understand and deploy, its predictive ability is limited, especially in the high COD range, where it is difficult to provide adequate support. This provides a baseline for subsequent models, but it is difficult to characterize complex pollution fluctuations accurately.

3.3 Lasso

3.3.1 Model Training

Like the MLR model, the dataset is divided into a training set and a testing set, with a ratio of 8:2. During the model training process, the optimal value of the regularization strength parameter α is selected through cross-validation and optimized based on the evaluation criterion of minimizing the mean absolute error (MAE). The model only trains variables that the Z-score has standardized to eliminate the influence of dimensional differences on the fitting process.

3.3.2 Prediction Results

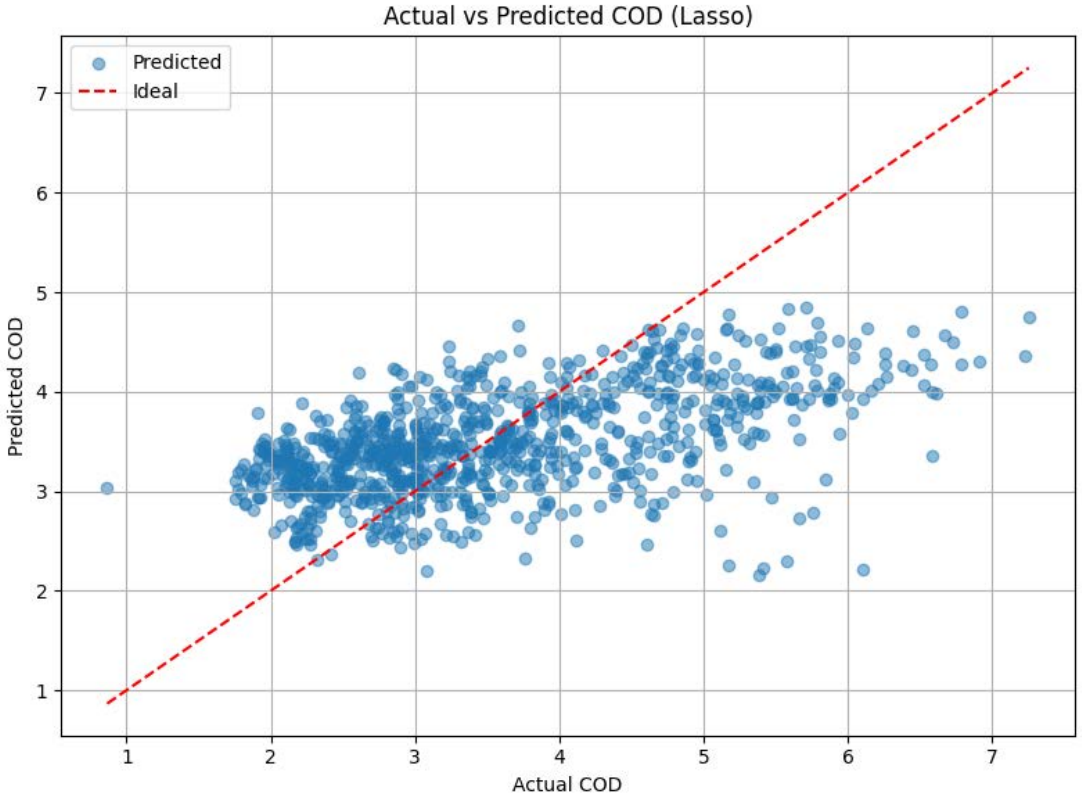

Fig. 4 Lasso model prediction result chart

ISSN 2959-6157

Data from: National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform

Picture credit: Original

Figure 4 shows the relationship between the predicted values of the Lasso regression model on the test set and the actual COD concentration. In the mainstream range of COD concentration between 2.7-3.4 mg/L, the samples are dense, and the predicted points are roughly distributed near the ideal fitting line, indicating that the model has a specific fitting ability in this concentration range. However, it can also be observed in the figure that the model deviates significantly in the high COD value range (>4mg/L), and the predicted values are generally lower than the actual values, indicating that the model's response to the increase in pollution concentration has a certain lag.

In addition, in the range of low COD concentration (<2.5 mg/L), some predicted results are significantly concentrated, basically higher than the fitted line, lacking differences, and may not accurately reflect the weak fluctuations in pollution concentration. This phenomenon suggests that the model tends to compress when processing edge concentration samples, affecting its sensitivity to outliers and rare cases.

The Lasso model can provide relatively stable prediction trends in the mainstream concentration range of COD but performs poorly in the extreme range, with a large distance between the fitted lines and actual fluctuations. This performance limits its predictive effectiveness in responding to high pollution events and may make it more suitable for daily monitoring.

3.3.3 Model Evaluation

The overall performance of the Lasso regression model on the test set is weak, with the following evaluation metrics:

R^2: 0.294. The model can only explain about 29.4% of the variation in COD concentration.

MAE: 0.778. The average deviation between each predicted value and the actual value is close to 0.78 mg/L.

RMSE: 0.978. Indicating significant fluctuations and extreme deviations in the error.

Combined with Fig.4, these indicators reflect that although the model performs well in the low to medium-concentration range, it shows significant disadvantages when dealing with high-concentration COD samples. The overall fitting ability is limited, and the error distribution structure tends towards the characteristics of "upward pressure" and "high concentration diffusion" in the predicted results.

The fitting equation of this model is as follows:

$$COD = 3.446 + 0.392pH + 0.229TN + 0.183TP + 0.001Temp \quad (2)$$

From the regression equation, the Lasso model retains the four variables of pH, TN, TP, and Temp and compresses the coefficient of NH_4 to 0, which is not included in the regression expression, reflecting the sparsity feature of the model in variable selection. Among them, TN and pH have the highest weights in the model, with coefficients of 0.229 and 0.392, respectively, indicating a close relationship with COD concentration. TP also shows a certain degree of positive correlation, while temperature variables have a weaker impact. The overall value of the coefficient is relatively small, which affects the model's ability to express COD changes.

Overall, the Lasso model sacrifices a certain level of prediction accuracy while implementing feature compression. Although its structure is relatively simple, its ability to handle complex changes in COD concentration is limited and insufficient to meet the needs of high-precision prediction.

3.4 Decision Tree Regressor

3.3.1 Model Training

The decision tree regression model uses a nonlinear structure based on conditional judgment rules. In this study, the maximum depth is controlled at 4 (max_depth=4) to avoid overfitting and improve the interpretability and operational efficiency of the model. This model does not rely on parameter fitting but is based on layer-by-layer partitioning of data features for prediction, which is suitable for scenarios with complex relationships between variables or significant nonlinear features.

This study used the Python platform to construct a decision tree model, with a data partitioning ratio of 80% for the training set and 20% for the testing set. Due to its clear model structure, shallow hierarchy, and fast inference speed, decision tree regression has good lightweight characteristics in resource-constrained scenarios.

3.3.2 Prediction Results

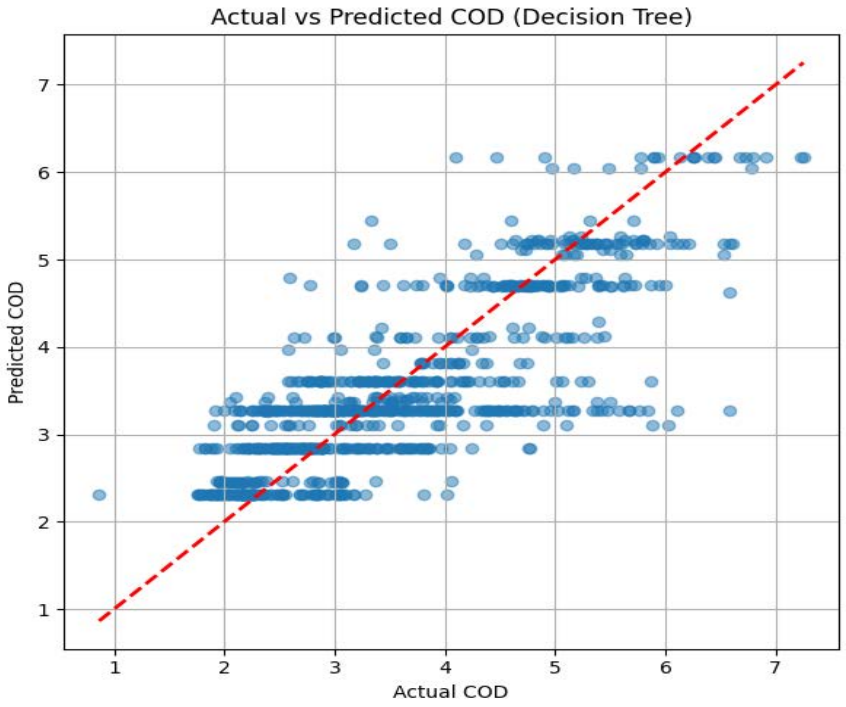

Fig. 5 Decision Tree model prediction result chart

Data from: National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform

Picture credit: Original

Figure 5 shows the correspondence between the predicted COD concentration by the decision tree regression model on the test set and the actual observed values. From the figure, the overall distribution of predicted values is relatively concentrated, especially in the mainstream range of COD concentration of 3-5 mg/L. The predicted results are close to the actual values, showing a certain degree of fitting ability [8].

In addition, in areas with high (>6 mg/L) or low (<2 mg/L) COD concentrations, the prediction error has expanded. Concentrated prediction and repeated values in some sample values indicate that the model's expressive ability in the edge region is limited. Based on the image results, the model performs robustly in the mainstream sample range but lacks fine-fitting ability in the extreme value range, indicating specific issues with its explanatory and generalization abilities.

Although the shallow tree structure set in this study limited model complexity and improved interpretability and execution efficiency, it also resulted in output values that were not smooth enough and lacked continuous transition features. This applies to the lightweight modeling requirements in resource-constrained scenarios, but it implies limitations in dealing with high-complexity pollution fluctuations.

3.3.3 Model Evaluation

The decision tree regression model performs better on the test set than the two linear models mentioned above, with the following evaluation metrics:

R^2: 0.631. The model can explain approximately 63.1% of the variation in COD concentration.

MAE: 0.519. The average deviation between each predicted value and the actual value is 0.519 mg/L.

RMSE: 0.707. Indicating relatively few significant extreme deviation values in error.

Figure 5 shows the correspondence between the predicted results and actual values of the decision tree model on the test set. The model performs relatively accurately in the low concentration range of COD, and the scatter points are distributed around the ideal prediction line with a reasonable degree of fitting. However, in the high concentration range, especially in areas where the COD concentration is higher than 6 mg/L, there are biases and fluctuations in the predicted results, reflecting the weak generalization ability of the model in sparse sample areas. In addition, several "horizontally stacked" concentrated predicted values were in the prediction results, indicating that the model tends to output discontinuities, consistent with the essence of segmented prediction in decision tree

ISSN 2959-6157

structures.

The decision tree model has good explanatory and predictive abilities, a transparent structure, and flexible variable processing. It is suitable for lightweight modeling scenarios with limited data scale and high modeling efficiency requirements in this study. Although there is still some error in the high concentration range, its overall fitting performance is significantly better than linear models, which can provide more reliable support for analyzing pollution concentration trends.

3.5 Auto Regressive Model

This study further introduces a time series model based on historical data. Based on the first three regression models, which are constructed based on the cross-sectional relationship between variables, this study aims to more fully capture the dynamic trend of COD concentration in the time dimension. After initial attempts at moving average (MA) and auto regressive (AR) models, it was found that the AR model performed better in fitting and stability. Therefore, the ARIMA (0,0,1) model, a simple auto regressive model, was ultimately chosen.

The model takes standardized COD data as input variables and uses the complete time series collected earlier (a total of 4755 items) for modeling. The evaluation results show that the model's RMSE on the test set is 0.707, which is better than the first three regression models. This indicates that it has a stronger fitting ability for the temporal fluctuation characteristics of COD.

The estimated results of the model are as follows: the constant term is 3.4586, and the coefficient of the moving average term (ma. L1) is 0.7837, both of which have passed the 1% significance level test, further indicating a significant auto correlation relationship between the current COD value and the previous period value and are suitable for short-term trend prediction. In addition, from the results of the Ljung Box test and Jarque-Bera test of residuals, the model conforms to the assumptions regarding randomness and normality, and the residual sequence has no significant auto correlation, which meets the requirements of subsequent analysis.

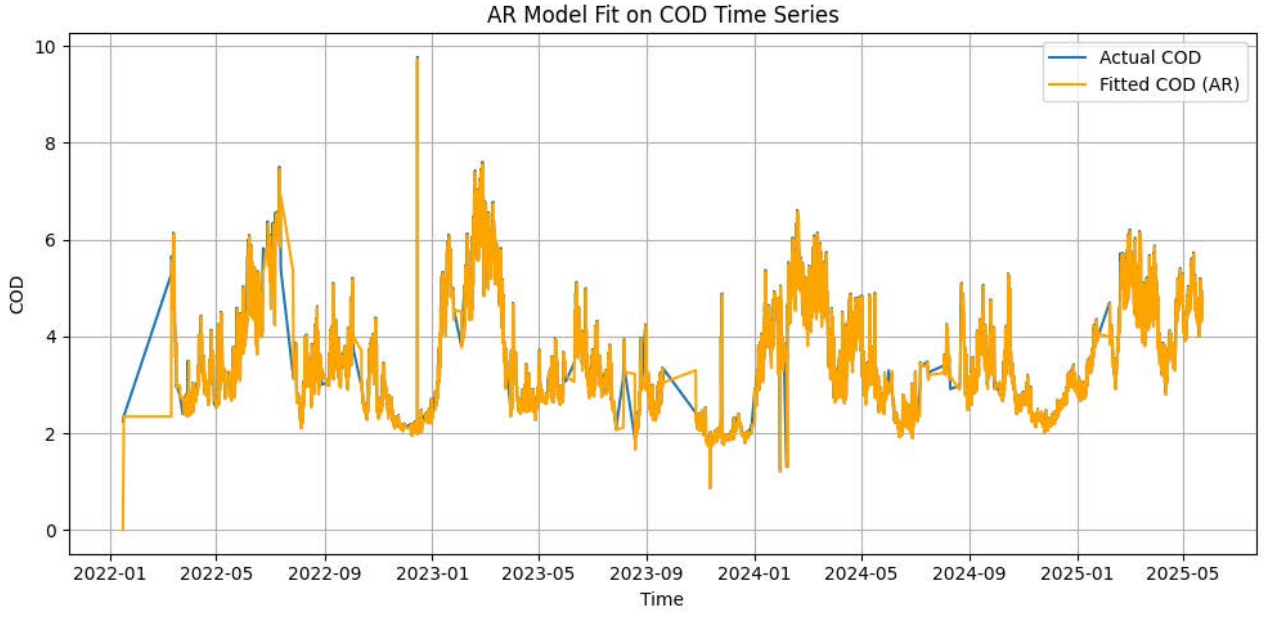

Fig. 6 AR Model Fit on COD Times Series

Data from: National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform

Picture credit: Original

From the results in Figure 6, the AR model fits the time variation process of COD concentration well in the overall trend. The curve is synchronized with the observed values at multiple peaks and valleys, especially from mid to late 2023 to early 2024. The model can stably capture seasonal fluctuations and sudden upward trends, indicating its strong practicality and fast response ability in processing continuous water quality data of this type. However, it can also be observed that there is a certain deviation in specific periods (such as early 2022 and around May 2024), and there is still room for improvement in extreme fluctuations or sparse sample intervals.

Although the AR model does not involve other environmental variables, its excellent time-fitting performance

provides an important supplement for COD concentration prediction within this study's time scale. In the future, the prediction results will be compared and explained in conjunction with the regression model.

4. Result

4.1 Model Comparison

4.1.1 Comparison of prediction performance between regression model and time series model

To predict COD concentration, this study constructed three types of regression models, namely multiple linear regression (MLR), LASSO regression, and decision tree regression, as well as one type of time series model (autoregressive AR), and obtained their respective evaluation indicators. Based on this, the performance of the models was systematically compared, providing modeling support for subsequent pollution identification and influencing factor analysis.

Regarding regression models, the MLR model fits each input variable through the least squares method. Although the model structure is simple and the computational cost is low, its performance in the face of nonlinear pollution changes is minimal. The R^2 on the test set is only 0.399, and the RMSE is 0.903, indicating that the model can only explain 39.9% of COD concentration changes. At the same time, the prediction bias is significant in high-concentration areas, and there is obvious underfitting, which cannot achieve good prediction within the allowable error range under the premise of limited data and resources.

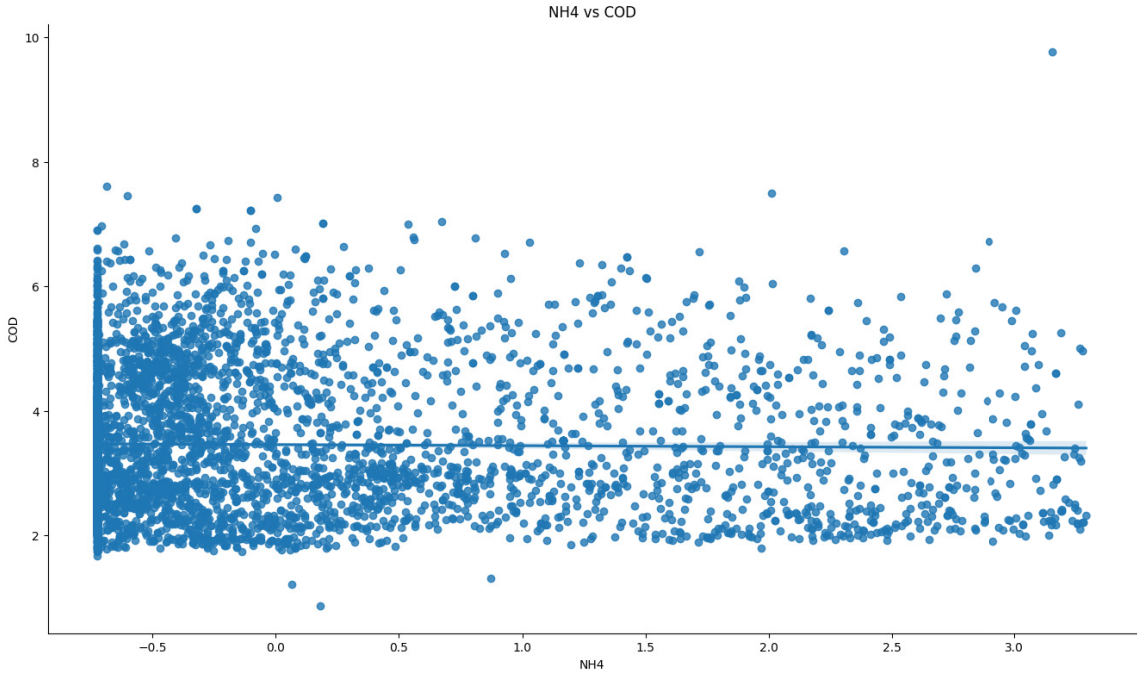

Fig. 7 Scatter Plots of NH4 vs. COD

Data from: National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform

Picture credit: Original

The LASSO model processes feature variables by applying compression constraints, automatically removing ammonia nitrogen variables and not including them in the final prediction equation while retaining the remaining four input factors. This screening result is not accidental. From the scatter plot of NH_4 and COD, it can be observed that there is no evident linear trend between the two. The data points show a horizontal band distribution and do not show significant changes in different concentration intervals. This indicates that the correlation between the NH_4 variable and COD is weak, or its influencing mechanism is not linearly interpretable. Therefore, ammonia nitrogen is considered a redundant or noisy variable in the Lasso screening process, and its removal has a certain rationality.

However, Lasso indirectly led to a decrease in fitting ability while reducing the complexity of the model. Compared with the MLR model, the R^2 (coefficient of determination) of the LASSO model further decreased to 0.294, and the RMSE (root mean square error) increased to 0.978, indicating that the fitting effect of the LASSO model in this data structure is not as good as that of the MLR model.

ISSN 2959-6157

This may be due to LASSO's failure to effectively capture key relationships in the data during variable compression or insufficient correlation between variables to support effective compression [9].

The output results of the LASSO model also show the phenomenon of "centralized prediction and edge collapse," especially when dealing with extreme value samples; prediction accuracy significantly decreases. This may be because LASSO tends to retain variables closely related to the target variable when selecting variables while ignoring the complex patterns of change that may exist in extreme value samples, resulting in poor predictive performance of the model for extreme value samples [10].

In contrast, as a non-parametric learning method, decision tree models can segment fit nonlinear relationships and interactions between variables. In this study, to control model complexity and enhance interpretability, the maximum depth of the tree was set to 4 layers, ensuring that the model runs in a lightweight framework. The results showed that the model achieved R^2=0.631, MAE=0.519, and RMSE=0.707 on the test set and performed best among all regression models. From the prediction performance, the decision tree model fits well in the mainstream concentration range of COD, maintains a relatively stable deviation in high-concentration areas, and the error is randomly distributed without showing systematic drift. This indicates that the model has balanced fitting ability, expression accuracy, and noise resistance [11].

The analysis of feature importance shows that TN is the most important predictor in the model, with a feature weight of up to 36.4%; Water temperature and pH followed closely behind, accounting for 21.9% and 17.8%, respectively; TP ranks fourth, accounting for 14.6%. This indicates that TN performs stably in multiple linear regression and Lasso models and dominates the decision path in tree models. The ammonia nitrogen that is automatically compressed to zero in Lasso also shows the lowest importance in the decision tree, accounting for only 9.3%, which verifies its limited linear explanatory power for COD concentration from another perspective.

However, all three regression models are based on the structure of "predicting target variables with input variables," which to some extent relies on the multidimensional integrity of the data and the collaborative changes between variables. When external influencing factors are missing, data volatility increases or monitoring frequency decreases, the adaptability and stability of regression models will be limited. To overcome this problem, an autoregressive time series model was introduced in the study, which uses "self-historical values" to predict the future trend of COD concentration, thereby reducing dependence on the quality and structure of input variables.

The AR model can still achieve an RMSE (0.707) equivalent to that of a decision tree without the involvement of external variables and can fit pollution fluctuations well over continuous periods. Compared with regression models, AR models show better sensitivity to temporal trends and can reflect the rising or falling pollution trend earlier.

Based on the analysis of comprehensive prediction accuracy and model applicability, both the AR model and the decision tree model have strong predictive performance. However, the former has more advantages in dynamic trend capture and deployment simplicity. Therefore, subsequent analysis will be based on the AR model to extrapolate pollution trends and identify potential influencing factors.

4.1.2 AR model prediction analysis for the next 7 days

Figure 8 shows the 7-day (168-hour) trend prediction of COD concentration based on the AR model. This model only uses historical COD concentration data for modeling and extrapolation, and the prediction expression is as follows:

$$COD_t = 3.4586 + 0.9951 \times COD_{t-1} \quad (3)$$

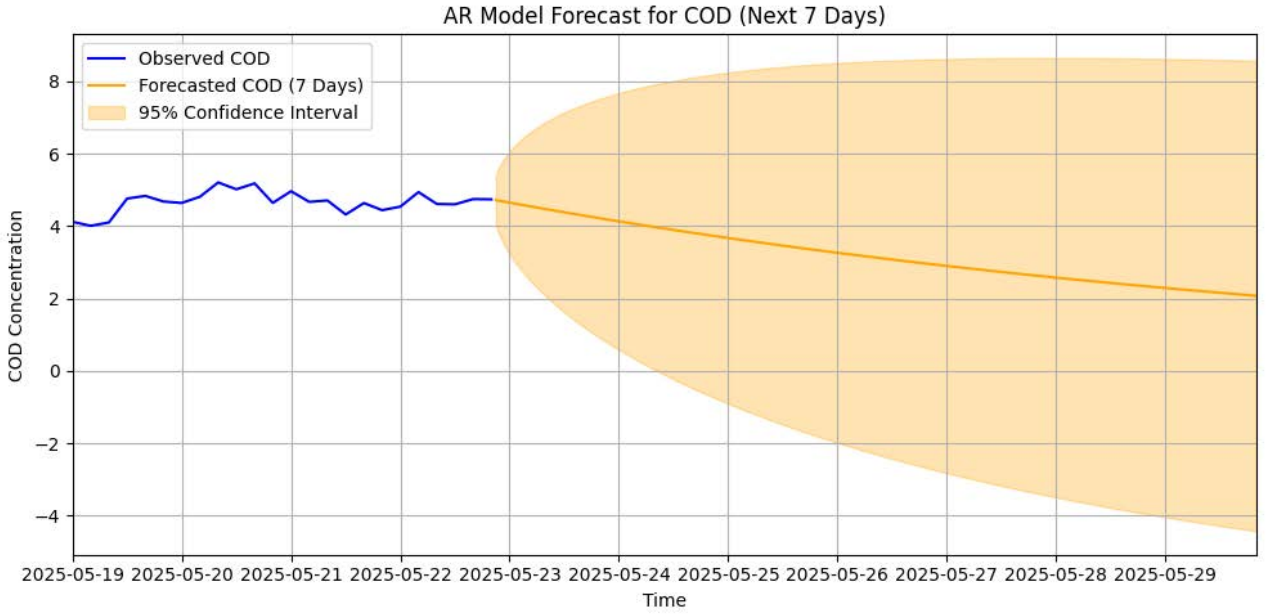

Fig. 8 AR model prediction (for the next 168 hours)

Data from: National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform

Picture credit:Original

Among them, the first-order lag coefficient is close to 1, indicating that the COD sequence has significant autocorrelation, and the concentration change in the short term depends on its previous state. From the predicted curve, the COD concentration will show a slight fluctuation and an overall stable, slight decrease trend in the next week, which is in line with the natural evolution law of regional water quality under non-flood season conditions. The orange line in the figure represents the predicted value, and the orange shaded area represents the 95% confidence interval. It gradually expands over time, showing a typical "fan-shaped" trend, reflecting the uncertainty of the model in long-term prediction.

To verify the effectiveness of the prediction, this article introduces 7 days of actual observation data from May 19th to 25th, 2025 and calculates the daily average COD concentration separately. The measured daily average is as follows:

May 19th: 4.415 mg/L; May 20th: 4.914; May 21st: 4.621; May 22nd: 4.694; May 23rd: 4.544; May 24th: 4.845; May 25th: 4.907.

Comparing these measured data with the predicted results of the AR model, the two maintain a high consistency in overall trends and values. From May 21st to 23rd, the difference between the predicted and measured values was slight, indicating that the model can continue the primary trend. All measured values are within the predicted interval, indicating that the model has a reasonable confidence range and basic boundary control capability. Further quantitative evaluation was conducted to obtain the mean absolute error (MAE) and root mean square error (RMSE) between the predicted and measured values of the AR model, which were 0.183 mg/L and 0.210 mg/L, respectively. This error level is within the allowable range of typical surface water pollution monitoring accuracy, indicating that the model is reliable in fitting COD concentration in the short term [12].

From the perspective of error distribution, the most significant errors in the 7 days occurred on May 20th and May 24th. The measured COD values for these two days were slightly higher than the overall trend, while the model-predicted values were more stable, showing a phenomenon of "underestimating the peak." This deviation reflects the response lag of the AR model in dealing with local abnormal fluctuations, which is an inherent limitation of its univariate linear structure. The model only uses the previous period values for recursion, lacking the ability to perceive non-linear jumps or external sudden disturbances. Therefore, when faced with occasional increases in actual data, the predicted values tend to be conservative, which lowers accuracy. During relatively stable COD concentrations (such as May 21-23), the AR model's predicted results are highly consistent with the measured values, with an error controlled within ± 0.1 mg/L. This indicates that the AR model has good trend continuity in environments with small fluctuations and strong trend continuity and stability. This is especially suitable for early warning judgment of short-term water quality changes.

ISSN 2959-6157

Although the model has a certain degree of delayed response at extreme values or mutation points, its overall error level is low, and it exhibits a strong fitting force in the trend area. Considering its advantages of simple structure, low training requirements, and no need for additional feature variables, AR models can be used as effective prediction tools in lightweight monitoring scenarios. They are especially suitable for regional water monitoring tasks with continuous and controllable data fluctuations.

4.2 Analysis of influencing factors

Among the various regression models constructed in this study, the feature importance ranking of the decision tree model provides a quantitative basis for analyzing the main control factors of COD concentration changes. Specifically, total nitrogen has the highest importance among all input variables, at 0.364, followed by water temperature and pH, at 0.219 and 0.178, respectively. Total phosphorus and ammonia nitrogen have lower importance, at 0.146 and 0.093, respectively. Meanwhile, in the LASSO model, ammonia nitrogen was not included in the final equation due to its weight being compressed to zero, further confirming its insufficient contribution to the current data structure.

From the perspective of synergistic changes in pollutants, COD concentration reflects the organic pollution load that can be oxidized in water bodies, which includes both natural sources (such as humus) and human activity sources (such as domestic sewage and agricultural runoff) [13]. As a common nutrient, total nitrogen is widely present in domestic wastewater, agricultural leachate, and some industrial wastewater. Its source and distribution often overlap highly with COD.

In the Xianyang Weihe River Basin, the significant loss of nitrogen fertilizer has significantly increased the concentration of TN in the water. It may drive the increase of COD levels through various mechanisms. On the one hand, excessive application of nitrogen fertilizer enters water bodies in the form of runoff under rainfall or irrigation conditions, which not only increases the input of inorganic nitrogen but also carries crop residues and other organic matter, increasing the organic load in water bodies; On the other hand, the enrichment of nitrogen promotes microbial activity, and microorganisms consume a large amount of dissolved oxygen during the decomposition of organic matter, further pushing up COD. In addition, rural aquaculture and agricultural non-point source pollution are commonly present in the region, showing a superimposed trend, exacerbating the compound input of nitrogen and organic matter. Existing monitoring data shows that the annual COD emissions from agricultural sources in the Weihe River Basin reach 125700 tons, with ammonia nitrogen at 6700 tons, of which rural breeding and fertilizer application are considered the main sources. Therefore, while nitrogen fertilizer loss increases TN concentration, it is indeed accompanied by mechanistic organic matter input, which has a significant impact on COD concentration [14]. In other words, the statistical correlation between TN and COD reflects the consistency of their pollution source structures.

The impact mechanism of water temperature is more reflected in the regulation of ecological and chemical processes. Higher temperatures enhance microbial activity in water bodies, thereby accelerating the decomposition of organic matter. At the same time, it may also trigger sediment release, inhibit dissolved oxygen, and indirectly increase COD concentration [15]. Especially before and after the rainstorm, the disturbance of the water body increases, and the organic particles are easily resuspended, further stimulating the short-term fluctuation of COD. Water temperature can be a good input variable and reflect information on meteorological and hydrodynamic processes.

Although pH value is not a widely defined direct pollutant indicator, its changes have a significant regulatory effect on water environmental processes. On the one hand, pH affects the dissolution equilibrium of metal ions and nutrients, indirectly affecting the stability and availability of organic matter. On the other hand, pH also determines the structure and metabolic pathways of microbial communities in water bodies, thereby affecting the degradation efficiency of organic matter. Most microorganisms exhibit high activity in neutral to weakly alkaline environments, while their organic matter degradation rate may be limited under acidic or strongly alkaline conditions [16].

The measured data in this study area indicate that the pH value is stable primarily, around 8.0, in a relatively typical weakly alkaline environment. In this context, pH ranks third in importance in the model, reflecting its indirect regulatory effects on microbial activity and organic matter conversion processes. On the one hand, stable alkaline conditions may facilitate the continuous decomposition of organic matter, thereby affecting the dynamic changes in COD concentration. On the other hand, slight fluctuations in pH may also alter microbial community function, causing periodic fluctuations in degradation capacity. Therefore, although pH is not a direct source of pollution load, its performance in predictive models still reveals its potential role as a water quality regulating factor.

Total phosphorus and ammonia nitrogen did not show significant predictive ability in this dataset. TP often exists in granular form, and its mobility is poor in non-rainstorm seasons, so it is difficult to dominate COD changes quick-

ly. Although NH_4 is an important form of nitrogen, its correlation with organic pollution is not as strong as TN overall, and its numerical fluctuation is relatively high, which may be weakened in modeling due to its "noise" attribute. In both LASSO and decision tree models, there is a weak correlation or a tendency to be excluded, indicating that their role in the watershed and dataset is relatively marginal.

This study did not directly introduce hydrological and meteorological factors such as precipitation, flow velocity, and runoff intensity. However, water temperature indirectly carries information about seasonality and environmental disturbances. Previous studies have shown that COD concentration exhibits significant seasonal fluctuations, driven by multiple factors such as rainfall and climate [17]. Afterward, external variables such as meteorological data, land use types, and flow velocity changes can be incorporated into the modeling system to improve predictive capabilities further.

Overall, the variation of COD concentration is influenced by multiple factors such as pollution source structure (such as nitrogen and phosphorus load), physical environment (such as water temperature and pH), and seasonal disturbances. In this study's dataset and watershed context, total nitrogen, water temperature, and pH constitute the core explanatory variables and dominate model prediction.

5. Conclusion

In this paper, the applicability of three lightweight prediction models is proposed and verified to address the realistic challenges of discontinuous data acquisition and low frequency in COD monitoring. Through the systematic preprocessing and variable screening of the monitoring data of Weihe River Tieqiao section, the input variables of the model were reasonably streamlined. The multiple linear regression model is suitable as the baseline of the lightweight model due to its simple structure and low computational cost, but its prediction accuracy is insufficient in the extreme value interval, suggesting that future research should consider introducing a moderate nonlinear component to enhance the prediction ability. This paper shows that under the conditions of limited data and computational resources, relying on traditional statistical methods can also realize the effective prediction of COD short-term trend, which provides a feasible path for the continuity and timeliness of water quality monitoring, and has practical application value for improving the level of water environment management. In the future, the model performance can be further optimized by combining more time series information and improving the algorithm structure.

References

[1] Zhang G, Du Q, Lu X, et al. A Novel Hybrid Strategy for Detecting COD in Surface Water[J]. Applied Sciences, Multidisciplinary Digital Publishing Institute, 2020, 10(24): 8801–8801.Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.

[2] Heping L. Establishment of a Monitoring System of Eco-Environment – Suggestions and Reflection[J]. Proceedings of the 2021 5th International Seminar on Education, Management and Social Sciences (ISEMSS 2021), 2020.

[3] Çimen Mesutoğlu Ö, Gök O. Prediction of COD in industrial wastewater treatment plant using an artificial neural network[J]. Scientific Reports, Springer Science and Business Media LLC, 2024, 14(1).

[4] Jung W S, Kim S E, Kim Y D. Prediction of Surface Water Quality by Artificial Neural Network Model Using Probabilistic Weather Forecasting[J]. Water, 2021, 13(17): 2392.

[5] Cheng Q, Kim J-Y, Wang Y, et al. Novel Ensemble Learning Approach for Predicting COD and TN: Model Development and Implementation[J]. Water, MDPI AG, 2024, 16(11): 1561.

[6] Zare Abyaneh H. Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters[J]. Journal of Environmental Health Science and Engineering, 2014, 12(1).

[7] National Ministry of Environment. National Water Quality Automatic Comprehensive Supervision Platform [EB/OL]. Cnemc.cn. 2023. https://szzdjc.cnemc.cn:8070/GJZ/Business/Publish/Main.html.

[8] Islam M, Shehzad F. A Prediction Model Optimization Critiques through Centroid Clustering by Reducing the Sample Size, Integrating Statistical and Machine Learning Techniques for Wheat Productivity[J]. M. Rahimi. Scientifica, 2022, 2022: 1–11.

[9] Tibshirani R. Regression Shrinkage and Selection Via the Lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267–288.

[10] Meng F. Economic Forecasting with Many Predictors[EB/OL]. TRACE: Tennessee Research and Creative Exchange. 2017/2025-05-25. https://trace.tennessee.edu/utk_graddiss/4483.

[11] Loh W-Y. Classification and regression trees[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1(1): 14–23.

[12] Arfastya F D, Wulandari S Y, Rifai A. Studi Persebaran Kandungan Fosfat dan Material Padatan Tersuspensi di Perairan Muara Sungai Slamaran, Kota Pekalongan[J]. Journal of Marine Research, 2023, 12(4): 563–570.

[13] Sharma K, Singh N. Comparative Analysis of Different Methods used in Evaluation of Physio-chemical Parameters of Surface Water: A Review[J]. International Journal of Research in Advent Technology, MG Aricent Private Limited, 2023, 11(4):

ISSN 2959-6157

13–21.
[14] ZHAO Z. ANALYSIS OF REMOTE SENSING TECHNOLOGY APPLIED ON HYDROLOGY AND WATER RESOURCES –TAKING WEIHE'S ECOLOGY AS AN EXAMPLE[J]. Applied Ecology and Environmental Research, Hungarian University of Agriculture and Life Sciences, 2019, 17(5).
[15] Neidhardt H, Shao W. Impact of climate change-induced warming on groundwater temperatures and quality[J]. Applied water science, Springer Nature, 2023, 13(12).
[16] Ulezlo I V, Bezborodov A M. Consumption of volatile organic compounds by alcaliphilic microorganisms[J]. Applied Biochemistry and Microbiology, Pleiades Publishing, 2007, 43(2): 197–200.
[17] Zhou F, Lu X, Chen F, et al. Spatial-Monthly Variations and Influencing Factors of Dissolved Oxygen in Surface Water of Zhanjiang Bay, China[J]. Journal of Marine Science and Engineering, 2020, 8(6): 403.