Theory and Practice of Science and Technology
2024, VOL. 5, NO. 4, 30-35
DOI: 10.47297/taposatWSP2633-456904.20240504

Research on the Early Warning for Abnormal Inflow in Wastewater Treatment Plants

Yaorong Shu[1,2], Linghao Chen[2], Tianming Wu[3], Mengjie Song[3], Wen Zhang[3], Juan Mao[2*]

[1]Institute of Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, 430074, China

[2]School of Environmental Science and Engineering, Huazhong University of Science and Technology, Wuhan, 430074, China

[3]Yangtze Ecology and Environment Co., Ltd., Wuhan 430074, China

ABSTRACT

In response to issues such as high process load and potential risks of failing to meet discharge standards caused by sudden water pollution incidents in combined industrial and domestic wastewater treatment plants, this study focused on the wastewater treatment plant in an industrial park. The Autoregressive Integrated Moving Average (ARIMA) algorithm was utilized to construct an influent water quality prediction and anomaly warning model. Autocorrelation function (ACF), partial autocorrelation function (PACF), and augmented Dickey-Fuller (ADF) tests are employed to determine the optimal parameters for predicting chemical oxygen demand (COD), ammonia nitrogen (NH_3-N), total nitrogen (TN), and total phosphorus (TP), optimizing the water quality prediction models. The impact of different monitoring frequencies and prediction steps on the model's performance were studied. Furthermore, an anomaly warning model for water quality was developed using an autoencoder (AE) and K-nearest neighbors (KNN) algorithm. This work provides a scientific basis for wastewater treatment plants to proactively respond to sudden water pollution events and lays a theoretical foundation for the intelligent transformation of traditional wastewater treatment plants.

KEYWORDS

Wastewater Treatment Plant; Time Series Prediction; ARIMA; Abnormal Warning

Currently, there are over 9, 000 industrial parks across the country, and in many small and medium-sized industrial parks, industrial wastewater is often treated together with domestic sewage due to local economic constraints. These high-concentration, poorly biodegradable, and highly toxic industrial wastewaters far exceed the design capacity of municipal wastewater treatment plants, often leading to the collapse of biological treatment systems, impacting the stable operation. Therefore, developing new methods for early warning of influent anomalies and establishing accurate and timely warning systems to detect and respond to abnormal influent events is of great significance.

Currently, wastewater treatment plants determine whether influent water quality is abnormal based on online monitoring data, operational experience of process facilities, and design values of wastewater quality [1]. Typically, online monitoring includes physical monitoring, chemical monitoring, and biological monitoring [2-4]. Although physical monitoring methods respond quickly and are easy to maintain [5], they have low sensitivity and single indicators. Chemical monitoring methods are highly accurate and provide comprehensive water quality indicators, but they respond slowly and are complex to maintain [6]. Biological monitoring methods can comprehensively reflect the impact of water quality on biochemical systems. However, they respond slowly, have a lag, and the uncertainty caused by multiple factors can lead to false alarms [7, 8]。 Therefore, it is necessary to study efficient, accurate, and timely methods for early warning of influent water quality.

In recent years, machine learning algorithms such as Support Vector Machine (SVM) [9, 10], Decision Tree (DT) [11]、K-nearest Neighbors (KNN) [12], Random Forest (RF) [13], Long Short-Term Memory (LSTM) [13], Artificial Neural Networks (ANN) [14] and their various derivative models have been widely used in various environmental predictions. Zhou et al. [15] used the RF model to predict the influent flow of the Humber wastewater treatment plant. Minsoo [16] used the KNN algorithm to predict the influent flow, COD, suspended solids, total phosphorus, and total nitrogen of a wastewater treatment plant. Yi Man et al. [17] proposed a hybrid artificial intelligence model for predicting the influent COD load of a wastewater treatment plant, using a Vector Autoregression (VAR) model combined with an Autoregressive Moving Average (ARIMA) model. These AI-based water quality prediction models combined with anomaly detection classification algorithms, they provide new ideas for early warning of water quality.

In this study, continuous hourly water quality data were obtained through automatic sampling and offline detection. After comparing the performance of various algorithms, the model parameters were optimized. Furthermore, the impact of data with different frequencies was explored, and the accuracy of the prediction model under different prediction steps was analyzed. On this basis, the accuracy of different methods for identifying water quality anomalies was compared, providing scientific guidance for water quality early warning in sewage plants.

1 Data and Methods

1.1 Data collection

The data used in this study were sourced from on-site collection of influent and effluent water samples from a composite sewage treatment plant in Hubei Province. The sampling scheme involved installing automatic water samplers (JC-8000H, Juchuang Environ. Protection. Ltd.). The sampling period was from March 24, 2023, to April 18, 2023, totaling 25 days. The sampling frequency was once per hour, and the samples were preserved at 4°C. The main detection indicators were chemical oxygen demand (COD), ammonia nitrogen (NH_3-N), total nitrogen (TN), and total phosphorus (TP). A total of 576 sets of data and 2304 water quality indicators were collected.

1.2 Data pretreatment

Due to equipment failure on April 6th, some data were missing. To ensure the completeness of the data in the time dimension, linear interpolation was used to fill in the missing values. Therefore, in this study, data preprocessing was performed using the locally weighted scatterplot smoothing (Lowess, LS), the rolling window (RW), and the Savitzky-Golay (SG) methods.

1.3 Model construction

The model construction in this study includes both a prediction model and an anomaly warning model.

Besides, three different anomaly detection methods, Autoencoder (AE), KNN, and Confidence

Interval, are used to identify abnormal water quality data and construct the early warning model.

1.4 Model evaluation

This study evaluates the prediction performance using the coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE). The formulas are as follows [18, 19]:

$$R^2\left(y,\hat{y}\right) = 1 - \frac{\sum_{i=0}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{m}(y_i - \bar{y})^2} \tag{1}$$

$$\mathrm{MAE}\left(y,\hat{y}\right) = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right| \tag{2}$$

$$\mathrm{MAPE}\left(y,\hat{y}\right) = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{3}$$

$$\mathrm{RMSE}\left(y,\hat{y}\right) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2} \tag{4}$$

where y_i is the actual sample value, $\widehat{y_i}$ represents the predicted value, $\bar{y}$ is the sample mean, m is the number of samples.

The performance of the water quality anomaly detection model was evaluated using accuracy, precision, recall, and F_1 score, whose formula are as follows [20, 21]:

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\mathrm{F\_1\ Score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{8}$$

2 Results and Discussion

2.1 The impact of data preprocessing on prediction model

In the prediction of COD, compared to the raw data without preprocessing, the performance of the COD prediction model improved after using LS and SG preprocessing, except for the data processed by RW. For the prediction models of NH_3-N and TN, the performance of the prediction models showed similar conclusions to the COD prediction model. For the TP prediction model, although the MAPE of the prediction model decreased from 54.22% to 47.53% after RW preprocessing, the MAE remained unchanged. Compared to SG, LS smooths the data by performing nonlinear fitting locally, effectively avoiding errors caused by random factors while preserving the original local features. Therefore, LS was ultimately chosen as the data smoothing preprocessing method.

2.2 Prediction model construction and optimization

2.2.1 Model construction

To explore the best algorithm for constructing prediction models, four conventional time series prediction algorithms, including Autoregressive (AR), Moving Average (MA), ARIMA, and Prophet, were compared. For COD prediction, Prophet had the worst performance among the four algorithms. In contrast, the ARIMA-based COD prediction model achieved better results, with an R^2 of 0.64, MAE of 17.75, MAPE of 12.44%, and RMSE of 27.51. Similarly, Prophet also performed poorly in the prediction models for NH_3-N, TN, and TP, while ARIMA consistently showed the best performance. The ARIMA algorithm considers both historical data and historical errors in time series prediction [22], making it more suitable for continuous influent water quality changes. Therefore, considering historical data and errors, the ARIMA algorithm was ultimately chosen to construct the water quality prediction model.

2.2.2 Hyperparameter optimization

The main hyperparameters of the ARIMA model are the autoregressive order (p), the differencing order (d), and the moving average order (q). The ADF test is used to determine whether the data is a stationary series, and for non-stationary series, (d)-order differencing is performed. The values of (q) and (p) are calculated using a combination of ACF and PACF [23].

The ADF test results indicated that for COD, the critical values at 1%, 5%, and 10% significance levels are -3.44, -2.87, and -2.60, respectively. The computed t-test value was -2.85, which was lower than the 5% critical value. This suggested that the data used in the experiment was already stationary and did not require further differencing; therefore, d is set to 0. Additionally, the ACF results become truncated after the 2nd order, so q is set to 2. The PACF results indicate that after the 7th order, the PACF results become tailing off, so p is set to 7. Therefore, for the COD prediction model, the optimal hyperparameters are p=7, d=0, and q=2.

Similarly, the optimal ARIMA parameters were determined to be p=7, d=0, and q=2 for NH_3-N prediction, p=4, d=0, and q=6 for the TN prediction, and p=5, d=0, and q=2 for TP prediction.

2.3 Times series analysis

2.3.1 The impact of monitoring frequency

For the COD prediction model, the model's performance rapidly declines with increasing monitoring intervals. When the monitoring interval exceeds 4 hours, the R^2 drops below 0, indicating that the model no longer has predictive capability. For the NH_3-N prediction model, when the monitoring interval exceeds 6 hours, the R^2 drops below 0, making the model unsuitable for predicting NH_3-N. For the TN and TP prediction models, the impact of different data acquisition frequencies is similar to that of the COD prediction model. Specifically, for constructing the NH_3-N prediction model, the monitoring frequency should be maintained between 1 to 6 hours, while for COD, TN, and TP predictions, the monitoring frequency should be maintained between 1 to 4 hours.

2.3.2 The impact of prediction steps

In the COD prediction process, as the step size increases from 1 to 4, the model's error gradually increases. When the step size is between 4 and 6, the MAPE fluctuates around 10%, indicating a gradual decline in model performance. When the step size exceeds 6, the MAPE is greater than 10%, and both MAE and RMSE are greater than 10, suggesting that the model is no longer suitable for COD prediction. For NH_3-N and TP prediction, when the step size exceeds 3, the MAPE is greater than 20%, making the model unsuitable for NH_3-N and TP prediction. For the TN prediction model, the trend of error variation is similar to other prediction models. Therefore, in water quality prediction, ARIMA-based models can provide relatively accurate predictions for COD up to 6 hours ahead, NH_3-

N and TP up to 2 hours ahead, TN up to 1 hour ahead.

2.4 Influent Water quality anomaly warming

2.4.1 Anomaly Labeling based on the correlation of influent and effluent

To more reasonably and accurately identify anomalies, a method based on the correlation between influent and effluent water quality has been proposed. Pearson correlation analysis of influent and effluent data indicated that the effluent at time □ had a high correlation with the influent from t-12 to t-15. Additionally, the hydraulic retention time of the wastewater treatment plant is 12 hours. Therefore, using a 12-hour time sequence as a lead time, the influent times that cause the effluent to exceed the standard can be back-calculated and marked as anomalous influent. This method was used to further evaluate the effectiveness of anomaly classification.

2.4.2 Comparison of anomaly classification algorithms

For COD anomaly detection, AE and KNN algorithms exhibit better performance compared to the confidence interval method in terms of accuracy. However, compared to the confidence interval method, machine learning-based anomaly detection algorithms have lower stability, precision, recall, and F_1 score. Stability will significantly improve with an increase in anomalous data samples. For NH_3-N, TN, and TP anomaly detection, the overall trend is similar to COD anomaly detection. However, for NH_3-N and TP anomaly detection, the KNN algorithm demonstrates the best performance. Therefore, AE are suitable for early warning of COD and TN anomalies based on time-series prediction algorithms, while KNN are more suitable for t NH_3-N and TP.

3 Conclusions

In this study, an ARIMA time-series prediction model was developed and optimized using ACF, PACF, and ADF methods to accurately predict water quality indicators in a composite wastewater treatment plant. The impact of monitoring frequency on the model's performance revealed that higher frequency data enhances the construction of prediction models. Specifically, for NH_3-N prediction, monitoring frequency should be maintained between 1 to 6 hours, while for COD, TN, and TP predictions, it should be between 1 to 4 hours. The research on prediction horizons indicated that the ARIMA-based models can accurately predict COD for up to 6 hours, NH_3-N and TP for 2 hours, and TN for 1 hour. Furthermore, anomaly detection algorithms were employed to identify influent anomalies, with the AE-based anomaly detection for COD and TN anomalies, respectively, and the KNN algorithm for NH_3-N and TP anomalies, respectively. This approach enables early warning of sudden influent anomalies in composite wastewater treatment plants and provides scientific guidance for the intelligent and precise management of wastewater treatment facilities.

Funding

This research received from the Research Project Funded by Yangtze Ecological Environmental Protection Group Co. Ltd. (grant number HBHB2022018).

References

[1] NEWHART K B, HOLLOWAY R W, HERING A S, et al.. Data-driven performance analyses of wastewater treatment plants: A review. Water Res,2019,157:498.

[2] MARTíNEZ R, VELA N, EL AATIK A, et al..On the Use of an IoT Integrated System for Water Quality Monitoring and

Management in Wastewater Treatment Plants. Water,2020,12(4):1096.
[3] LEE M W, HONG S H, CHOI H, et al.. Real-time remote monitoring of small-scaled biological wastewater treatment plants by a multivariate statistical process control and neural network-based software sensors. Process Biochem, 2008,43(10):1107-1113.
[4] MELIDIS P, VAIOPOULOU E, AIVASIDIS A. Development and implementation of microbial sensors for efficient process control in wastewater treatment plants. Bioprocess Biosyst Eng,2008,31(3):277-282.
[5] SONG C, ZHANG H. Study on turbidity prediction method of reservoirs based on long short term memory neural network. Ecol Model,2020,432.
[6] CHING P M L, SO R H Y, MORCK T. Advances in soft sensors for wastewater treatment plants: A systematic review. J Water Process Eng,2021,44:102367.
[7] GUST M, BURONFOSSE T, GEFFARD O, et al.. In situ biomonitoring of freshwater quality using the New Zealand mudsnail Potamopyrgus antipodarum (Gray) exposed to waste water treatment plant (WWTP) effluent discharges. Water Res,2010,44(15):4517-4528.
[8] HYBSKá H, LOBOTKOVá M, VANEK M, et al.. Biomonitoring and its in the assessment of the quality of wastewater treatment process. Environ Nanotechnol, Monit Manag,2020,13:100292.
[9] WANG Q, LI Z, CAI J, et al.. Spatially adaptive machine learning models for predicting water quality in Hong Kong. J Hydrol,2023,622:129649.
[10] YANG Z, ZOU L, XIA J, et al.. Inner Dynamic Detection and Prediction of Water Quality Based on CEEMDAN and GA-SVM Models. Remote Sens-Basel,2022,14(7):1714.
[11] TAHRAOUI H, AMRANE A, BELHADJ A-E, et al.. Modeling the organic matter of water using the decision tree coupled with bootstrap aggregated and least-squares boosting. Environ Techno Inno,2022,27:102419.
[12] PESANTEZ J E, BERGLUND E Z, KAZA N. Smart meters data for modeling and forecasting water demand at the user-level. Environ Modell Softw,2020,125:104633.
[13] CHEN K, CHEN H, ZHOU C, et al.. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Res,2020, 171:115454.
[14] EGBUERI J C, AGBASI J C. Performances of MLR, RBF-NN, and MLP-NN in the evaluation and prediction of water resources quality for irrigation purposes under two modeling scenarios. Geocarto Int,2022,37(26):14399-14431.
[15] ZHOU P, LI Z, SNOWLING S, et al.. A random forest model for inflow prediction at wastewater treatment plants. Stoch Env Res Risk A,2019,33(10):1781-1792.
[16] KIM M, KIM Y, KIM H, et al.. Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. Front Env Sci Eng,2016,10(2):299-310.
[17] MAN Y, HU Y, REN J. Forecasting COD load in municipal sewage based on ARMA and VAR algorithms. Resour Conserv Recy,2019,144:56-64.
[18] LU H, MA X. Hybrid decision tree-based machine learning models for short-term water quality prediction. Chemosphere,2020,249:126169.
[19] SINGHA S, PASUPULETI S, SINGHA S S, et al.. Prediction of groundwater quality using efficient machine learning technique. Chemosphere,2021,276:130265.
[20] NASIR N, KANSAL A, ALSHALTONE O, et al.. Water quality classification using machine learning algorithms. Journal of Water Process Engineering,2022,48
[21] MA M, ZHAO G, HE B, et al.. XGBoost-based method for flash flood risk assessment. J Hydrol,2021,598:126382.
[22] WANG G, SU H, MO L, et al.. Forecasting of soil respiration time series via clustered ARIMA. Comput Electron Agr, 2024,225.
[23] BILGILI M, PINAR E. Gross electricity consumption forecasting using LSTM and SARIMA approaches: A case study of Türkiye. Energy,2023,284:128575.