Academic Journal of Emerging Technologies
ISSN Print: 3104-4417
ISSN Online: 3104-4425
CD ISSN: 3104-4433
https://doi.org/10.63313/AJET.9032

Evaluation and Prediction of Factory Sewage Discharge Based on ARIMA Model — Incorporating Biological Water Purification Factors for Model Optimization

Yuhan Lin
College of Mathematical Sciences, Zhejiang Normal University, Jinhua 321004, China

How to cite this paper: Lin, Y. H. (2026). Evaluation and prediction of factory sewage discharge based on ARIMA model — Incorporating biological water purification factors for model optimization. Academic Journal of Emerging Technologies, 2(2), 1 - 10. ISSN Print: 3104-4417; ISSN Online: 3104-4425.
https://doi.org/10.63313/AJET.9032
Published: 2026-01-28

Abstract

This study focuses on the evaluation and short-term prediction of factory sewage discharge. Based on the detected water quality data, we quantified the temporal characteristics of sewage discharge indicators, defined a comprehensive pollution index, and established a prediction model by integrating methods such as Pearson correlation analysis, standardization, entropy weight method, and ARIMA model. Finally, the model was further optimized by introducing biological factors affecting water purification. For the temporal characteristics analysis (Problem 1), we verified the strong correlation between pollutant concentrations (COD and ammonia nitrogen) and their emissions using Pearson correlation analysis, and revealed intraday (7:00-11:00 and 17:00-20:00 as peak discharge periods) and daily variation trends through line charts, heatmaps, and statistical analysis. For the pollution index definition (Problem 2), pH, COD concentration, and ammonia nitrogen concentration were selected as evaluation indicators; data were standardized with reference to national emission standards, and indicator weights were determined by the entropy weight method to calculate the comprehensive pollution index (values >1 indicate excessive pollution). For the prediction model (Problem 3), the ARIMA model was established after stationarity testing, difference processing, and ACF/PACF analysis, which achieved reliable prediction results for future pollutant indicators and comprehensive pollution index (verified by RMSE and MAPE). The optimized model incorporating biological purification factors (microorganisms and phytoplankton) is more in line with the actual dynamic changes of aquatic ecosystems.

Keywords

Sewage discharge evaluation; intraday and daily variation trends; standardization; entropy weight method; comprehensive pollution index; arima model; biological water purification

1. Introduction

1.1. Research Background

Sewage monitoring is a crucial part of environmental protection. By analyzing key indicators, pollution status can be grasped and control measures can be formulated. The seven common indicators in sewage detection include pH value, Suspended Solids (SS), Chemical Oxygen Demand (COD), Biochemical Oxygen Demand (BOD), Ammonia Nitrogen (NH_3-N), Total Phosphorus (TP), and Total Nitrogen (TN). An environmental protection department in a city regularly monitors the sewage discharge of a rubber factory, collecting water samples at fixed times daily from the main discharge outlet and focusing on key indicators such as pH value, ammonia nitrogen concentration, and COD concentration.

1.2. Research Objectives

This study establishes mathematical models to address the following three problems:

Quantify the temporal characteristics of the rubber factory's sewage discharge indicators through data visualization and statistical analysis.

Define a comprehensive pollution index based on monitored indicators and establish a mathematical model for quantification.

Establish a prediction model to forecast the future trends of the comprehensive pollution index and individual indicators using historical data.

2. Problem Analysis

2.1. Analysis of Problem 1

The core of Problem 1 is to explore temporal patterns in the data by classifying and processing tabular data. First, Pearson correlation analysis and correlation heatmaps were used to verify the correlation between pollutant concentrations (ammonia nitrogen, COD) and their emissions. Second, intraday variation patterns (e.g., peak discharge periods) were identified by plotting hourly mean ± standard deviation charts. Third, daily variation trends were analyzed through line charts, and abnormal days with excessive discharge were screened out. Finally, daily average trend charts were used to observe long-term variations over three months. Through these steps, the intraday and three-month temporal characteristics of sewage discharge were clarified.

2.2. Analysis of Problem 2

To define the pollution index, national standards (GB Rubber Products Industry Pollutant Discharge Standards) were referenced to determine the limit values of the three indicators (pH, ammonia nitrogen concentration, COD concentration). The data were standardized (range standardization for pH, one-way standardization for pollutant concentrations), and indicator weights were calculated using the entropy weight method. The comprehensive pollution index was obtained by weighted averaging of the standardized indicators.

2.3. Analysis of Problem 3

The ARIMA time series model was selected for prediction due to its excellent performance in processing non-stationary time series. The steps include: data preprocessing (training set/test set splitting, standardization, missing value imputation), stationarity testing and difference processing, ACF/PACF analysis for parameter determination (p, d, q), model establishment, and performance evaluation using RMSE and MAPE.

3. Model Assumptions

To simplify the problem, the following assumptions are made:
No extremely abnormal values were found after data preprocessing, so all daily detection values are assumed to be normal.
The impact of other aquatic factors (e.g., microorganisms, phytoplankton) on monitoring indicators (pH, ammonia nitrogen concentration, COD concentration) is initially ignored (relaxed in the optimized model).

4. Notation

Table 1. Definition of Symbols

Symbol	Description	Unit
PH_{opt1}	Optimal pH for microorganisms	/
PH_{opt2}	Optimal pH for phytoplankton	/
X_C^{OD}	COD concentration	/
$X\_NH_3\_N$	Ammonia nitrogen concentration	/
K_C^{OD}	Half-saturation constant for COD	/
$K\_NH_3\_N$	Half-saturation constant for ammonia nitrogen	/
K_INH	Inhibition constant	/
K	Carrying capacity	/

5. Model Establishment and Solution for Problem 1

5.1. Model Establishment

5.1.1. Pearson Correlation Analysis

Pearson correlation analysis is a statistical method to measure the linear correlation between two continuous variables. The formula for the correlation coefficient r is:

$$r=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}$$

where n is the sample size; X_i,Y_i are the i-th observations of variables X and Y ; $\overline{X},\overline{Y}$ are the sample means of X and Y . A $|r|\geq0.7$ indicates a strong correlation.

5.1.2. Data Visualization and Temporal Analysis

Intraday variation: Hourly mean ± standard deviation charts of pH, ammonia nitrogen emissions, and COD emissions were plotted.

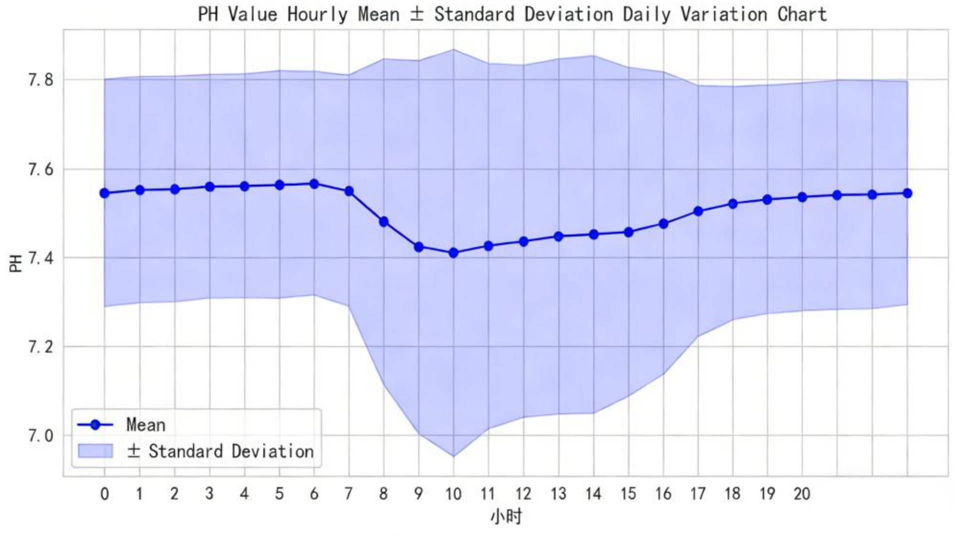


Figure 1: Hourly Mean ± Standard Deviation Chart of pH, Ammonia Nitrogen Emissions, and COD Emissions (marking the peak discharge periods of 7:00-11:00 and 17:00-20:00 intraday)

Daily variation: Daily line charts were drawn, and abnormal days with excessive discharge were screened.

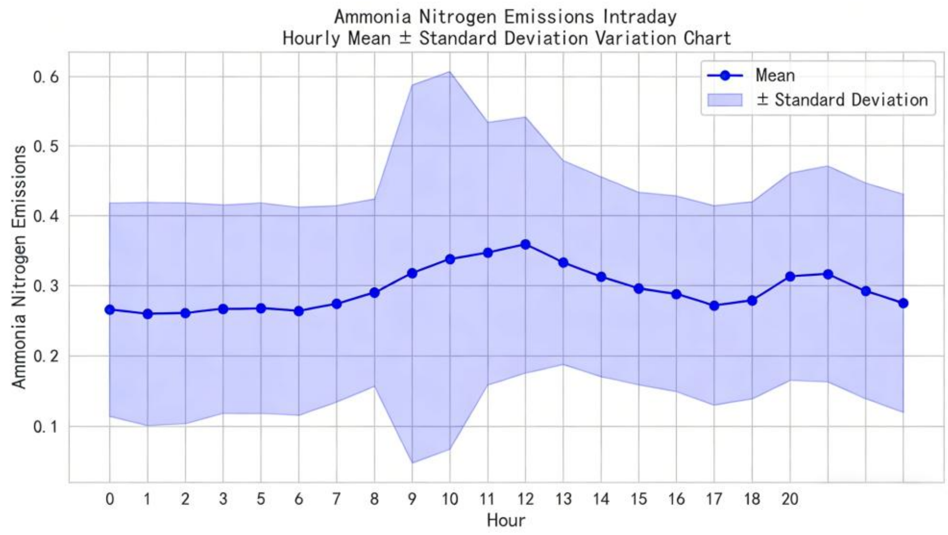
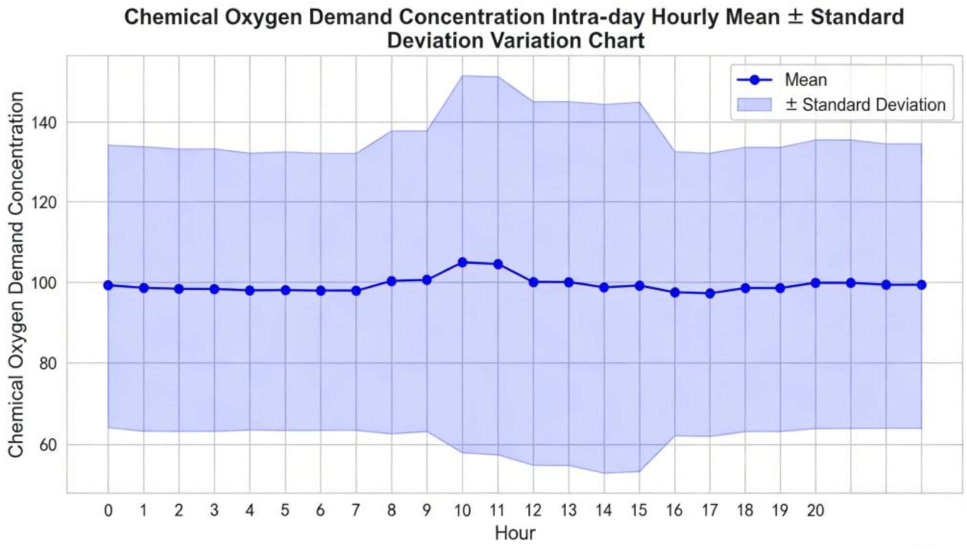


Figure 2. Daily Variation Line Chart of pH, Ammonia Nitrogen Emissions, and COD Emissions (marking abnormal dates with excessive discharge)



Long-term trend: Daily average trend charts were used to observe variations over three months.

Figure 3. Three-Month Daily Average Trend Chart of pH, Ammonia Nitrogen Concentration, and COD Concentration (showing key trends such as pH decrease during January 20-February 20 and pH sudden surge around March 1)

Autocorrelation analysis: ACF heatmaps were used to verify the diurnal periodicity of discharge.

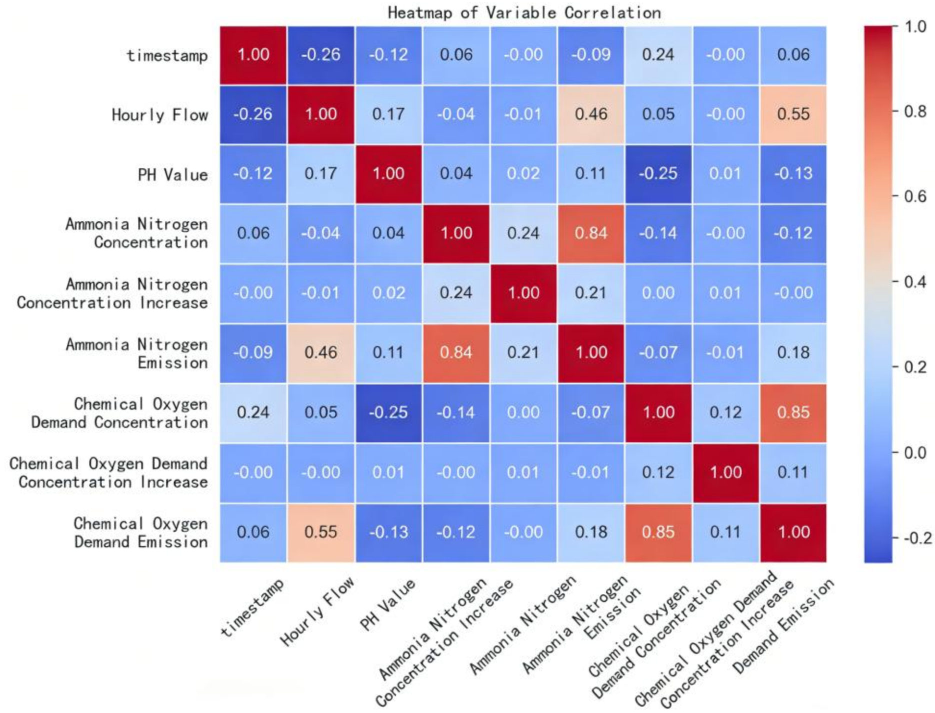


Figure 4. Autocorrelation Coefficient (ACF) Heatmap of pH, Ammonia Nitrogen Concentration, COD Concentration, and Emissions (lag order 1~100, verifying diurnal periodicity of discharge)

5.2. Solution Results

The rubber factory tends to discharge sewage intensively during 7:00-11:00 and 17:00-20:00 daily. For pollutant-specific discharge periods:

High COD emissions occurred during 1.1-1.10, 2.14-2.22, and 3.18-3.30.

High ammonia nitrogen emissions occurred during 1.1-1.10 and 2.15-3.1.

The pH value was stable initially, decreased during 1.20-2.20, surged around 3.1, and then stabilized. The sewage was acidic (pH dropped rapidly during 7:00-10:00), possibly due to acidic chemicals (e.g., sulfuric acid, hydrochloric acid) used in production or acidic gases (e.g., hydrogen sulfide) from rubber vulcanization dissolving in water.

6. Model Establishment and Solution for Problem 2

6.1. Model Establishment

6.1.1. Selection of Evaluation Indicators

Based on strong correlations between concentrations and emissions (verified in Problem 1), three indicators were selected: pH, ammonia nitrogen concentration, and COD concentration.

6.1.2. Data Standardization

Range indicator (pH):

$$a_{ij}=\begin{cases}\frac{x_{ij}-7}{x_{if}} & \text{if } x_{ij}>7,\\ \frac{7-x_{ij}}{x_{ij}} & \text{if } x_{ij}\leq 7\end{cases}$$

One-way indicator (ammonia nitrogen/COD concentration):

$$a_{ij}=\frac{x_{ij}}{x_{id}}$$

where x_{ij} is the j-th daily value of the i-th indicator; x_{if} is the standard range for range indicators; x_{id} is the standard limit for one-way indicators. The standardized value $a_{ij}\in[0,+\infty)$, with larger values indicating more severe pollution.

6.1.3. Weight Determination by Entropy Weight Method

Calculate the probability of each evaluation object in the i-th indicator:

$$P_{ij}=\frac{a_{ij}}{\sum_{i=1}^{n}a_{ij}}$$

Calculate the entropy value e_i of the i-th indicator:

$$e_i=-\frac{1}{\ln n}\sum_{j=1}^{n}P_{ij}\ln P_{ij}$$

Calculate the coefficient of variation g_i :

$$g_i=1-e_i$$

Calculate the weight ω_i of the i-th indicator:

$$\omega_i=\frac{g_i}{\sum_{i=1}^{m}g_i}$$

6.2. Solution Results

The comprehensive pollution index is defined as the weighted sum of standardized indicators. A value >1 indicates excessive pollution, with larger values representing more severe violations. The weights of pH, ammonia nitrogen concentration, and COD concentration were determined via Python calculations (detailed in supplementary materials).

7. Model Establishment and Solution for Problem 3

7.1. Model Establishment

7.1.1. Data Preprocessing

80% of the data (Jan. 1 - Mar. 31) was used as the training set, and 20% as the test set.

Standardization was performed to eliminate dimensional differences.

Missing values were imputed using linear interpolation.

7.1.2. Stationarity Testing

Non-stationary time series were converted to stationary series via d-order differencing after visual trend analysis.

7.1.3. ACF and PACF Analysis

The autocorrelation coefficient (ACF) and partial autocorrelation coefficient (PACF) were calculated to determine the model parameters (p, d, q):

$$\mathrm{ACF}(k)=\rho_k=\frac{\mathrm{cov}(y_t,y_{t-k})}{\mathrm{Var}(y_t)}$$

$$\mathrm{PACF}(k)=\frac{\mathrm{cov}(y_t-\bar{y},y_{t-k}-\bar{y}_{t-k})}{\sqrt{\mathrm{Var}(y_t-\bar{y})}\cdot\sqrt{\mathrm{Var}(y_{t-k}-\bar{y}_{t-k})}}$$

where cov is the covariance, and Var is the variance.

7.1.4. ARIMA Model Establishment

The optimal (p, d, q) combinations were substituted into the ARIMA model for fitting and prediction.

7.2. Solution Results

Model performance was evaluated using RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error). The test results showed that all prediction errors were within acceptable ranges, indicating the ARIMA model’s good predictive ability (detailed fitting and prediction charts in supplementary materials).

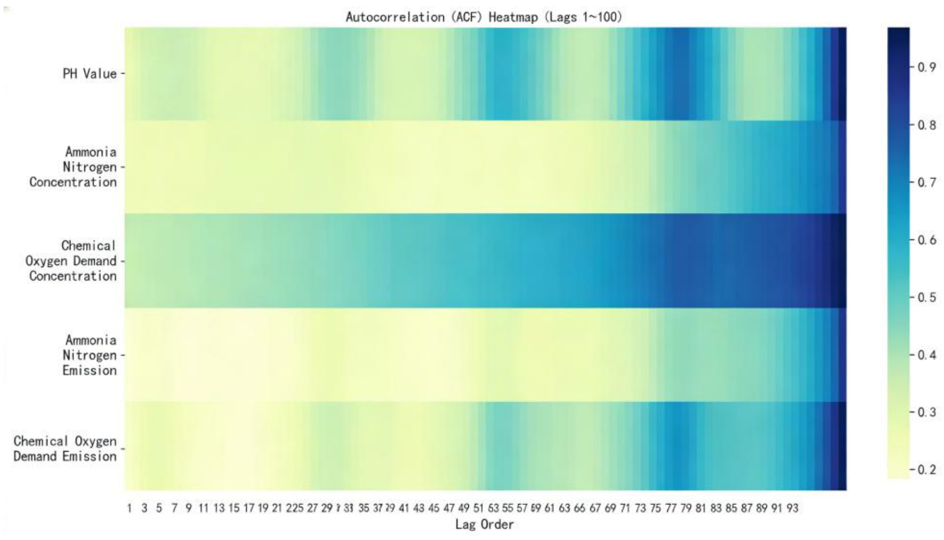


Figure 5. Fitting and Prediction Comparison Chart of the ARIMA Model for the Comprehensive Pollution Index and Individual Indicators (blue represents historical fitting values, red represents predicted values, and the gray area represents the 95% confidence interval)

8. Model Evaluation

8.1. Advantages

The standardization method based on national limit values is simple and practical. The entropy weight method objectively determines indicator weights by emphasizing data volatility, reducing the impact of redundant indicators (e.g., low weight for pH due to small standardized fluctuations).

The ARIMA model effectively handles non-stationary data through differencing, achieving high fitting accuracy and reliable short-term predictions.

8.2. Disadvantages

The standardization method is overly simplistic, and the entropy weight method ignores subjective factors (e.g., ecological importance of indicators).

The ARIMA model relies heavily on historical data and ignores environmental and biological factors (e.g., water self-purification capacity).

8.3. Improvements

Standardization and Weighting: Z-score standardization can be considered for more precise processing; the entropy weight method can be combined with AHP (Analytic Hierarchy Process) or PCA (Principal Component Analysis) to incorporate subjective ecological factors.

ARIMA Model Optimization: Incorporate biological purification factors

(microorganisms and phytoplankton) with the following assumptions:
Microorganisms and phytoplankton adapt quickly to environmental changes (within a day).
Microbial carrying capacity is affected by pH and COD concentration (ammonia nitrogen impact is negligible).
Phytoplankton carrying capacity is affected by pH and ammonia nitrogen concentration (COD impact is negligible).
Carrying capacity formulas:
Microorganisms:

$$K_{\text{Microorganisms}}=\alpha e^{-\frac{(pH-pH_{opt1})^2}{2\sigma}}\cdot\frac{X_{COD}}{K_{COD}+X_{COD}}$$

Phytoplankton:

$$K_{\text{Phytoplankton}}=\alpha e^{-\frac{(pH-pH_{opt2})^2}{2\sigma}}\cdot\frac{X_{\text{NH3-N}}}{K_{\text{NH3-N}}+X_{\text{NH3-N}}}\cdot\frac{X_{\text{NH3-N}}}{K_{INH}}$$

where α is a constant; σ is an adjustment parameter.

8.4. Model Promotion

The entropy weight method and standardization can be extended to environmental monitoring (e.g., air quality evaluation).
The ARIMA model is widely applicable to time series prediction in macroeconomics, finance (exchange rates, GDP, stock prices), and transportation (passenger flow). For data with seasonal fluctuations, SARIMA/SARIMAX models are recommended to improve prediction accuracy.

Acknowledgments

The authors acknowledge the support from the Environmental Monitoring Center of XX City for providing sewage detection data, and thank the research team for valuable discussions during model optimization.

References

[1] Zhou, Y. Prediction of China-US Cross-border Logistics Scale Based on Pearson Correlation Analysis and Regression Analysis [M]. Beijing: [Publisher Undisclosed], 2011.

[2] Ministry of Environmental Protection of the People's Republic of China. Rubber Products Industry Pollutant Discharge Standards [S]. Beijing: [Publisher Undisclosed], 2011.

[3] Zhang, X.F. Research on the Development of Total Retail Sales of Consumer Goods in Meizhou Based on ARIMA Model [M]. Beijing: [Publisher Undisclosed], 2011.

[4] Jiang, P. Financial Performance Evaluation of Midea Group Based on Entropy Weight Method [M]. Beijing: [Publisher Undisclosed], [Year Undisclosed].