Environment and Climate Protection (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/envcp.2025.040103
ISSN 2523-6504 Vol. 4 Num. 1

Study on Air Quality Index Forecasting in Nanjing Based on Time Series Modeling

Chenglong Chao, Huanzheng Zhu, Jiaqiang Xie, Zhengxun Fang

School of Mechanical and Electronic Engineering, Shandong Jianzhu University, Jinan, 250101, China

Keywords: Air Quality Index, GM(1, 1) Model, LSTM Model, ARIMA Model

Abstract: With the rapid advancement of industrialization and urbanization, air quality has become a global concern. In this paper, GM(1, 1) model, ARIMA model and LSTM model are used to predict the future air quality index in Nanjing. The GM(1, 1) prediction model takes the development coefficient = 0.00012, and the grey role quantity is 1435.236; the LSTM prediction model uses the mean square error (MSE) as the loss function, the Adam optimizer is optimized, the hidden nodes of the hidden layer are taken as 15, and the bath-size is taken as 1. The learning rate is 0.001, and the number of iterations is 300 times, and the ARIMA. The autoregressive order p of the prediction model is taken as 13; the difference order d is taken as 1; and the moving average order is taken as 2. Then the corresponding fitting effects are plotted according to the real and predicted values. Finally, by comparing RMSE, MAE and MAPE, it is concluded that the ARIMA model has better prediction effect. The selection of a suitable prediction model for the future AQI in Nanjing can provide more accurate AQI prediction for Nanjing, which is of great significance for promoting the green and sustainable development of Nanjing.

1. Introduction

With the rapid advancement of global industrialization and urbanization and the increasing frequency of human socio-economic activities, air quality issues have become a focus of attention worldwide. The deterioration of air quality not only affects the human living environment, but also poses a serious threat to people's health. Therefore, it is of great practical significance and urgent need to accurately predict the air quality index (AQI) to provide a scientific basis for the formulation and implementation of environmental protection policies.

Xu Faming et al [1] used gray system theory to construct a GM (1, 1) prediction model using Lushi County ambient air automatic monitoring data as a sample, and analyzed and predicted the trend of ambient air quality changes in the county; Wenqin [2] used the LSTM model to establish a prediction system, and called the prediction model based on the LSTM algorithm to predict the future short-term air quality index through the J2EE platform; Wang Jianshu et al [3] proposed an autoregressive integral sliding average (ARIMA) model-based prediction method for the air quality index of Suzhou City, using R software to process the daily air quality index data of Suzhou City in 2018, screening the best ARIMA model parameters, and then using the model to predict the air quality index in the first six days of January 2019.

The purpose of this paper is to forecast the AQI of Nanjing using the GM(1, 1) model, ARIMA model and LSTM model, and to compare the forecasting accuracy and effect of the three models. Specifically, this paper will construct three prediction models based on historical AQI data and set the corresponding parameters. Among them, the development coefficients and gray effect sizes of the GM(1, 1) model will be obtained based on data fitting; the LSTM model will use the mean squared error (MSE) as the loss function, and the Adam optimizer will be used for optimization with a reasonable number of hidden nodes, batch-size, learning rate, and number of iterations; and the ARIMA model will achieve the accuracy and effectiveness of the forecasting of Nanjing AQI through the selection of autoregressive order, difference order, and moving average order, which will be used to predict the AQI of Nanjing. ARIMA model, on the other hand, realizes the best fit to the data through the selection of autoregressive order, difference order and moving average order. After the model construction is completed, this paper will draw the fitting effect graphs under the three prediction models according to the real and predicted values to visualize the prediction performance of each model. Finally, the prediction accuracies of the three models are quantitatively evaluated by calculating the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) of each model, in order to determine the most suitable model to be used for the prediction of future AQIs in Nanjing. This study not only helps to improve the accuracy and reliability of air quality prediction, but also provides a scientific basis for the formulation of environmental protection policies, which has important theoretical and practical value.

2. Fundamentals of the model

The data in this paper are the open-source data collected by the China Air Quality Online Monitoring and Analyzing Platform, which is available at the following URL: https://www.aqistudy.cn/. The collected data include the AQI of Nanjing from 2018 to 2023, and the data from 2018 to 2022 are selected for prediction, because the AQI predicted in this paper is time series data, so the data from 2018 to 2022 are used as the model training set, and the data from 2023 are used as the test set.

2.1 GM(1, 1) model

2.1.1 GM(1, 1) model principle

GM(1, 1) model refers to the first order univariate gray prediction model. The “first order” is an equation that represents the uncertainties in the system by a definite mathematical relationship, i.e., the “whitened differential equation” is of the first order, and “gray” in the model refers to the “gray” system that contains partially known information and partially unknown information. The “gray” in the model refers to the “gray” system that contains part of the known information and part of the unknown information [4].

First, define the gray derivative of a sequence of numbers, namely:

$$d(k) = x^{(0)}(k) = x^{(1)}k - x^{(1)}(k-1) \tag{1}$$

Let $z^{(1)}(k)$ be the neighboring value generating series of the series $x^{(1)}$,namely:

$$z^{(1)}(k) = \alpha x^{(1)}(k) + (1-\alpha)x^{(1)}(k-1) \tag{2}$$

Thus the gray micro equation model of GM(1,1) is defined, namely:

$$d(k) + \alpha z^{(1)}(k) = b \text{ or } x^{(0)}(k) + \alpha z^{(1)}(k) = b \tag{3}$$

where $x^{(0)}(k)$ is called the gray derivative, α is called the development coefficient, and b is

called the gray role quantity.

Bring in the values of the moments and introduce the matrix vector a few numbers

$$x^{(0)}k = \begin{bmatrix}\alpha\\ b\end{bmatrix} \quad Y = \begin{bmatrix}x^{(0)}(1)\\ x^{(0)}(2)\\ \dots\\ x^{(0)}(k)\end{bmatrix} \quad B = \begin{bmatrix}-z^{(1)}(1) & 1\\ -z^{(1)}(2) & 1\\ \dots & \dots\\ -z^{(1)}(k) & 1\end{bmatrix} \tag{4}$$

This then leads to the GM(1,1) model, denoted as $WY = Bu$.

The estimate for solving the sum using the least squares method is:

$$u = \begin{bmatrix}\alpha\\ b\end{bmatrix} = (B^T B)^{-1} B^T Y \tag{5}$$

It is useful to treat the moment k in the gray differential equation as a continuous variable t. Then the previous series $x^{(1)}$ becomes a function of the variable t. Thus the gray derivative $x^{(0)}(k)$ corresponds to the derivative of the continuous function $\frac{dx^{(t)}}{dt}$, and the neighborhood generating series $z^{(1)}(k)$ corresponds to $x^{(1)}(t)$. Thus the gray differential equation for GM(1, 1) corresponds to the white differential equation:

$$\frac{dx^{(1)}(t)}{dt} + \alpha x^{(1)}(t) = b \tag{6}$$

Eq. (6) is solved to obtain the corresponding sequence of time for the GM(1,1) model, and then the prediction of the original data series can be obtained by cumulation.

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k), k = 1,2,\dots n, \tag{7}$$

2.1.2 Tests of the GM(1, 1) model

In order to ensure the feasibility of the GM(1, 1) modeling approach, it is necessary to test the degree of fit of the GM(1, 1) model to the oiginal data, which is tested in this section using two methods: residual test and posterior difference test method.

2.1.2.1 Residual test

From the above equation, the predicted value of the original data series can be calculated, and then the absolute residuals Δ_k, relative residuals $\varepsilon_r(k)$, and the average relative residuals $\overline{\varepsilon_r}$.

Judging Criteria: When the average relative residuals are used as a criterion to evaluate the effectiveness of the GM(1, 1) model, when $\overline{\varepsilon_r} < 20\%$, it indicates that when the GM(1, 1) model is utilized to fit the original data, the fitting effect obtained is acceptable. When $\overline{\varepsilon_r} < 10\%$, it is considered that the fit obtained when this prediction is fitted to the original data results using the GM(1, 1) model is good.

2.1.2.2 Posterior difference test

From the parameters of the above equations, the corresponding stage deviation and average stage deviation $\bar{\eta}$.

Judging Criteria: When the average level deviation is used as a criterion for evaluating the effectiveness of the GM(1, 1) model, when $\bar{\eta} < 0.2$, it indicates that when the GM(1, 1) model is utilized to fit the original data, the fitting effect obtained is acceptable. When $\bar{\eta} < 0.1$, it is considered that the fit obtained when this prediction is fitted to the original data results using the GM(1, 1) model is good.

2.1.3 Prediction of AQI based on the GM(1, 1) model

In this study, the data of Nanjing from January 2017 to December 2022 are used as the training set of the GM(1, 1) model, and the data from January to December 2023 are used as the test set, and the SPSS software is used to predict the air quality values of Nanjing from January to December 2023, and the development coefficient $\alpha = 0.00012$ is calculated, and the gray role quantity b is 1435.236, and the mean level deviation $\bar{\eta} = 12.809\%$, then the time response series of the GM(1, 1) model is:

$$\hat{x}^{(1)}(k+1) = (116 - \frac{1435.236}{0.00012})e^{-0.00012k}) + \frac{1435.236}{0.00012}, k = 1,2,\dots n, \quad (8)$$

Obtaining the true and predicted values along with their absolute errors and error rates as shown in Table.1.

Table.1. GM(1, 1) model AQI prediction data

Dates	Real value	Projected value	Absolute error	Inaccuracy
2023-01	70	73.78	3.78	5.40%
2023-02	58	73.69	15.69	27.06%
2023-03	75	73.60	1.40	1.86%
2023-04	88	73.51	14.49	16.46%
2023-05	82	73.42	8.58	10.46%
2023-06	95	73.33	21.67	22.81%
2023-07	58	73.24	15.24	26.28%
2023-08	86	73.15	12.85	14.94%
2023-09	67	73.06	6.06	9.05%
2023-10	77	72.97	4.03	5.23%
2023-11	63	72.88	9.88	15.69%
2023-12	77	72.79	4.21	5.46%

Analyzing the above table, it can be seen that the maximum error rate reaches 27.06%, and the error rate in July also reaches 26.28%, the error rate in three months is more than 20%, and the error rate in seven months is more than 10%, and the prediction results of using the GM(1, 1) model have relatively large errors, and the average relative residuals are calculated to be about 0.141 and the standard deviation is about 10.84, which is not a high degree of accuracy. The average error rate is 13.39%, and the fitting is good. The specific fitting effect is shown in the Figure.1.below.

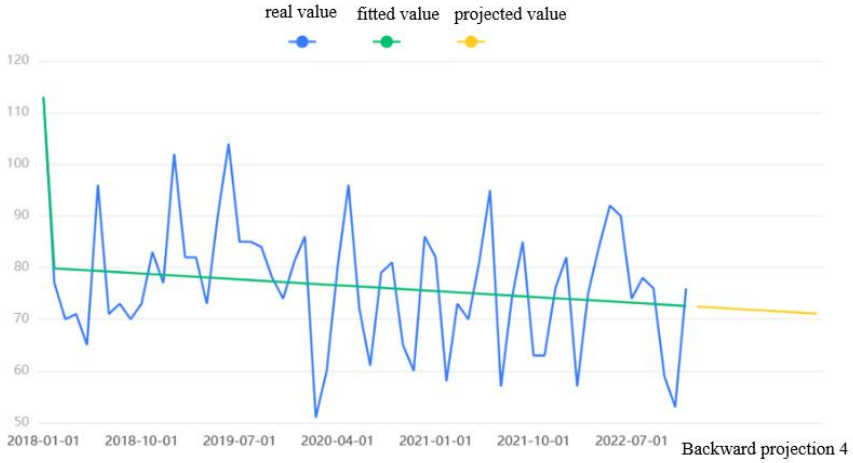

Figure.1. GM(1, 1) model fitting effect plot

From the fitting effect graph, the predictive fitting effect of GM(1, 1) model shows a decreasing state, the overall simulation effect is good, but the predicted value deviates from the real value by a large margin, the prediction accuracy is not high, and it is necessary to improve the GM(1, 1) model by processing the data of the original data as well as by correcting the residuals.

2.2 LSTM model

2.2.1 LSTM model principle

Long Short-Term Memory Network (LSTM, Long Short-Term Memory) is a special kind of Recurrent Neural Network (RNN) commonly used as a deep learning model for processing sequence data, especially in the field of Natural Language Processing (NLP) and time series prediction. It utilizes a unique gate control to achieve effective processing of sequence data, overcoming the problem of gradient vanishing or explosion of traditional RNNs on long sequences [5].

LSTM, proposed by Hochreiter and Schmidhuber in 1997, is able to efficiently capture and memorize long-term dependencies by introducing structural units called "gates". There are three gates inside the LSTM unit: a forgetting gate, an input gate and an output gate. Each gate has a learnable sigmoid activation function that controls the inflow and outflow of information.

The forgetting gate can be described as a filter that determines what information needs to be discarded from memory in order to maintain the long-term memory of the network.

The input gate determines what information needs to be added to memory from the current input, that is, what information needs to be memorized and stored. This consists of two parts: an update control part, which decides which quantities the main body of a book will update in the memory cell, and a hyperbolic tangent (tanh) layer, which creates a new vector of candidate values that passes through a sigmoid activation function to be selected before being added to the memory cell.

The decisions of the forgetting gate and the input gate are used to update the memory cells. Multiplying the cell state with the value of the forgetting gate indicates that we have forgotten part of the state information; then multiplying the value of the input gate with the candidate value and adding it indicates that we have added part of the new state information. The updated cell state can be obtained as Eq:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \tag{9}$$

where f_t is the output of the oblivion gate, C_{t-1} is the cell state at the previous moment, i_t is the input gate value, and $\tilde{C}_t$ is the candidate value.

The output gate determines the information to be output from the current memory. The current input and the hidden state from the previous time step are passed through a fully connected layer and a sigmoid function is applied to get the value of the output gate, which is between 0 and 1, with the closer to 0 indicating the smaller amount of information allowed to be output from the cell state, and the closer to 1 indicating the larger amount of information allowed to be output from the cell state. Then in this paper, the cell state is passed through the tanh function to get a value between -1 and 1, and multiplied with the value of the output gate to get the final hidden state.

Output gate formula:

$$o_t = \sigma(W_o \cdot [\hbar_{t-1}, x_t] + b_o) \tag{10}$$

$$\hbar_t = o_t * tan\hbar\,(C_t) \tag{11}$$

where W_o is the weight matrix, b_o is the bias term, h_{t-1} is the hidden state at the previous moment, x_t is the current input, W_o is the sigmoid function, C_t is the current cell state, and tanh is the hyperbolic tangent function.

For the LSTM model prediction analysis, the final output value h_t calculated by Eq. (11) is only a part of the completion of the analysis process, and it is also necessary to optimize the simulation effect by calculating the weight gradient and adjusting the weights to make the value of the loss function decrease continuously, so as to make the model prediction more accurate.

2.2.2 AQI prediction based on LSTM modeling

In this section, an LSTM neural network model is built with the PyTorch library in python to predict the air quality index AQI, eight typical AQI influencing factors are selected, the data is normalized, the data from January 2 018 to December 2022 is used as the training set, and the data from January 2023 to December 2023 is used as the test set, and the data is converted to PyTorch tensor for use in the PyTorch model, the loss function was used as Mean Square Error (MSE), and the optimizer was used as Adam. The model was iterated using the training data, the loss was computed and the parameters were updated. The performance of the model is evaluated using the test data and the loss values on the test set are printed, multiple predictions are made and finally the step size is determined to be 3, hidden nodes in the hidden layer are taken to be 15, the bath-size is taken to be 1, the learning rate is taken to be 0.001, and 300 iterations are made. Finally, the predictions of the model on the test set are back-normalized and written to an Excel file. Final comparison results, as shown in Table.2.

Table.2. LSTM model AQI prediction data

Dates	Real value	Projected value	Absolute error	Inaccuracy
2023-01	70	72.45	2.45	3.50%
2023-02	58	61.39	3.39	5.84%
2023-03	75	69.96	5.04	6.72%
2023-04	88	103.44	15.44	17.55%
2023-05	82	79.60	2.40	2.92%
2023-06	95	61.45	33.55	35.31%
2023-07	58	72.83	14.83	25.76%
2023-08	86	63.34	22.66	26.35%
2023-09	67	70.26	3.26	4.87%
2023-10	77	68.73	8.27	10.14%
2023-11	63	67.06	4.06	6.45%
2023-12	77	82.90	5.90	7.67%

As shown by the fitting effect graph, the prediction errors for July and August 2022 are large at 25.76% and 26.35%, respectively, which are more than 25%, and the error rate for June reaches 35.31%, and there are seven months where the error rate is less than 10%, with a good model accuracy, and an average error rate of 12.76% calculated. The fitting effect is shown in Figure 2.

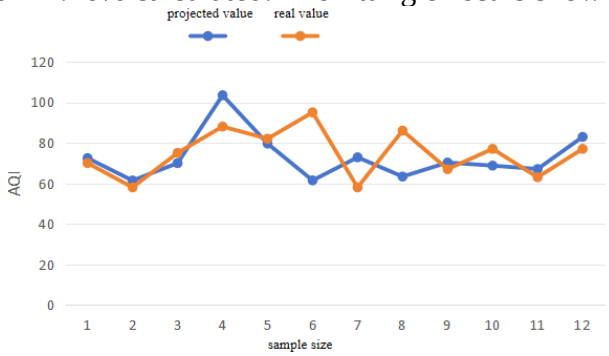

Figure.2. LSTM model fitting effect graph

As can be seen from the figure, the overall change is not very large, relatively smooth, especially in April, June, July and August of the predicted value deviation is larger, the other months of the predicted value of the basic and the real value of the deviation is not large, the total trend is similar to the real trend, the fitting effect is good!

2.3 ARIMA model

2.3.1 ARIMA model principle

The ARIMA model, as a typical model in the field of statistics, is also known as the autoregressive integrated moving average model, which can predict the future values by utilizing the past values of the time series. In the expression ARIMA(p, d, q), MA is the "sliding average", AR is the "autoregressive"; p is the autoregressive order; d is the difference order; q is the moving average order.

2.3.1.1 AR model

Represents the relationship between current observations and past observations. The AR part is denoted by p and is called the order. Specifically, the AR(p) model uses a linear combination of p past observations to predict the current value. The mathematical model expression is given below:

$$y_t = \mu + \sum_{i=1}^{p} r_i y_{t-i} + \varepsilon_t \tag{12}$$

where y_t is the current value, μ is the constant term, p is the order, r_i is the autocorrelation coefficient, ε_t is the random error term, and ε_t simultaneously conforms to a normal distribution.

The model reflects a linear relationship that exists between the target value at moment t and the target value before te first p moments t-1, t-2...t-p.

2.3.1.2 MA model

The moving average model is concerned with the accumulation of the error term in the autoregressive model, and the mathematical model expression is as follows:

$$y_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{13}$$

where θ_i is the moving average coefficient; and ε_{t-i} is the random error term at moment $t-i$. The model reflects a linear relationship that exists between the target value at moment t and the first t-1-p error values.

2.3.1.3 ARMA model

The model describes a combination of autoregression and moving average with the following mathematical modeling:

$$y_t = \mu + \sum_{i=1}^{p} r_i y_{t-i} + \varepsilon_i + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{14}$$

2.3.2 AQI forecasting based on ARIMA modeling

When fitting time series data using ARIMA model, it is first necessary to determine the respective orders of the autoregressive term (AR), the difference term (I), and the moving average term (MA),

and in the process of determining the orders, the graphs of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) are used in this section, and the initial determination of the model's order is made by observing the graphs' change in a certain lagged order, and utilizing the Python library to carry out the computation of ACF and PACF, and the obtained images are shown in Figure.3.

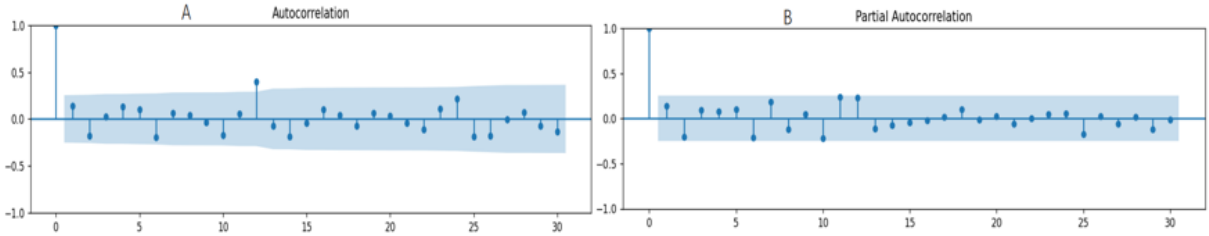

Figure.3. ACF Figure A and PACF Figure B

In the ACF plot, the horizontal coordinate is the lag order (lags) and the vertical coordinate is the autocorrelation coefficient. It can be observed that there is significant autocorrelation at positions of lags (lag) of 1 and 2, so one can try to select either 1 or 2 moving average terms. In the PACF plot, it can be observed that autocorrelation truncates after lags (lags) of about 2, so one can try to select 2 autoregressive terms. Initially, (2, 1, 2) was chosen as the order of the ARIMA model. After several fitting prediction calculations, (13, 1, 2) was finally selected as the final model order.

Table.3. ARIMA model AQI prediction data

Dates	Real value	Projected value	Absolute error	Inaccuracy
2023-01	70	76.78	6.78	9.69%
2023-02	58	58.3	0.3	0.52%
2023-03	75	70.19	4.81	6.41%
2023-04	88	80.03	7.97	9.06%
2023-05	82	85.11	3.11	3.79%
2023-06	95	81.72	13.28	13.98%
2023-07	58	76.61	18.61	32.09%
2023-08	86	82.7	3.3	3.84%
2023-09	67	80.19	13.19	19.69%
2023-10	77	71.78	5.22	6.78%
2023-11	63	68.39	5.39	8.56%
2023-12	77	82.90	5.90	7.67%

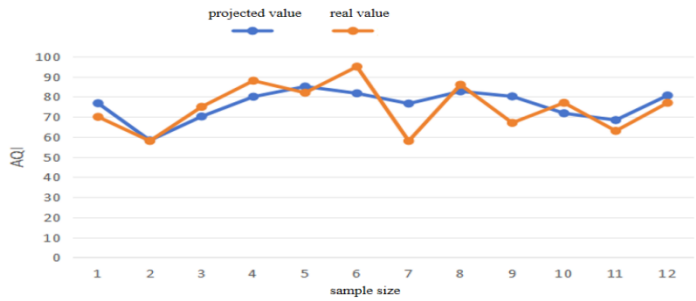

Figure.4. ARIMA model fitting results

As can be seen from Table 3 and Figure 4, the overall prediction effect based on the ARIMA model is good, with an average error rate of 9.93%, which improves the accuracy compared with the

previous models, except for the large deviation of the prediction value in July, the prediction error rate of other months is also basically controlled within 20%, and the error rate of nine months is below 10%, so the overall prediction knot is still good.

3. Results

After the models are built, in order to better compare the performance of each model and accurately predict the air quality conditions in Nanjing, this subsection compares them using three indicators: RMSE MAE MAPE.

RMSE (Root Mean Square Error) is a common measure of model prediction error and is often used to assess the performance of regression models. It calculates the average difference between the model's predicted values and the true observed values, i.e., the square root of the mean of the squares of the residuals. The smaller the root mean square error, the more accurate the model predictions.

MAE (Mean Absolute Error) is another common measure of model prediction error and is also commonly used to assess the performance of regression models. Similar to RMSE, MAE measures the difference between the predicted value and the true observed value, but it is the average of the absolute values of the calculated residuals.

MAPE (Mean Absolute Percentage Error) is a commonly used measure of prediction error and is particularly applicable to the assessment of the prediction accuracy of regression models at different scales. MAPE represents the mean percentage error between the predicted and true values. The smaller its value, the higher the predictive accuracy of the model.

The calculation of the three indicators was done in python and the results are shown in Table.4.

Table.4. Comparison of model prediction effects

Model	RMSE	MAE	MAPE
GM(1, 1)	11.479	9.823	13.392%
LSTM	13.773	10.104	12.790%
ARIMA	8.754	7.135	9.929%

Combined with the error analysis, the three indicators of the machine learning time series ARIMA model are smaller than the other two models, indicating that the model's performance is better, the accuracy is higher, and the prediction effect is more desirable, which can be used to predict the air quality index AQI of Nanjing, in which the MAPE of the LSTM is smaller than the GM(1, 1), which indicates that the accuracy of the LSTM is better in comparison with the other one, so the use of the machine learning ARIMA prediction model is more reliable.

4. Conclusions

In this paper, based on the air quality index data of Nanjing from 2018 to 2023, the GM(1, 1) model, the LSTM model and the ARIMA model are used to predict the air quality index AQI respectively, and analyze the advantages and disadvantages of the models, and the results show that the three models predict the future air quality of Nanjing, and the results obtained are all good, in which the ARIMA model has the highest prediction accuracy and is The ARIMA model has the highest prediction accuracy and is the most suitable for the prediction of future air quality index in Nanjing. The selection of a reliable and accurate air quality index prediction model can help to improve the prediction accuracy, help the relevant departments to make scientific and reasonable decisions, and create a good ecological environment for the development of the city.

The LSTM model has many hyperparameters to be adjusted, such as the size of the hidden layer, the learning rate, the bath-size, etc., which requires a lot of time and resources for tuning.The shortcomings of the ARIMA model lie in the lack of control over the error of the prediction based on

the temporal and multifactorial regression, and the question of whether the errors of the multiple predictions are in a controllable interval. In this paper the autoregressive order (p), the number of differences (d), and the moving average order (q) of the ARIMA model are determined by trial-and-error method, which is a relatively cumbersome process and results in large errors. To optimize the LSTM model, automated parameter tuning tools can be used for fast empirical-based tuning, and long-term dependency capture can be enhanced by stacking layers, adopting GRU or bi-directional structures, and incorporating attention mechanisms. Grid search, Bayesian optimization, or automated machine learning techniques can be used to reduce the difficulty of parameter selection for ARIMA models. In addition, multiple model combinations can be considered to fuse the performance of different models to better cope with the complex relationships and uncertainties in the data and thus improve the prediction accuracy of the AQI.

References

[1] Zhang H, Wang J, Nie Y.A novel optimization model based on fuzzy time series for short-term Air Quality Index forecasting[J].Knowledge-based systems, 2024(Jul.19):296.
[2] Chi P T L, Lan V T H, Hoang L T, et al. Impact of Burning Incense/Joss Paper on Outdoor Air Pollution: An Interrupted Time Series Analysis Using Hanoi Air Quality Data in 2020[J].Global Journal of Health Science, 2024, 16(3):8.DOI:10.5539/gjhs.v16n3p27.
[3] Wang Jianshu, Wang Ying, Zhao Minxian, et al. Application of ARIMA model in the prediction of air quality index in Suzhou City [J]. Public Health and Preventive Medicine, 2019, 30(02): 18-20.
[4] Li Jin, Yang Li, Chen Tiange. Prediction of ambient air quality in Anshan City based on gray GM(1, 1) model[J]. Qinghai Environment, 2023, 33(03): 147-153.
[5] Xia Yulu. A review of the development of recurrent neural networks[J]. Computer Knowledge and Technology, 2019, 15(21): 182-184.