AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj3017

PREDICTION OF AIR QUALITY INDEX BASED ON ARIMA AND LSTM MODELS—TAKE GUILIN AS AN EXAMPLE

Dan Chen
Department of Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.
Corresponding Email: 1274356593@qq.com

Abstract: With the acceleration of urbanization and the improvement of industrialization, air quality has become an important factor affecting the quality of life and health of urban residents. Under this background, the air quality problem has become increasingly prominent and has become a key factor affecting the quality of life and health of urban residents. Therefore, how to accurately predict the air quality index (AQI) in order to take timely and effective measures to protect the health of residents has become a major challenge we face. This study takes Guilin city as the object, and constructs an air quality prediction model by integrating big data and artificial intelligence technology. Based on the historical monitoring data such as PM2.5 and NO2, a training set/test set division method is adopted, and the ARIMA time series model and the LSTM depth learning network are comprehensively applied: the former captures the trend and seasonal characteristics of the data, and the latter handles the complex sequence relationship. By introducing the rolling prediction mechanism to update the model parameters in real time, the dynamic adaptability of the prediction system is effectively improved. Experiments show that the hybrid model significantly improves the prediction accuracy and stability of AQI, and provides a scalable technical scheme for urban air quality management.
Keywords: LSTM model; ARIMA model; Rolling forecast; Air quality index forecast

1 INTRODUCTION

1.1 Research Background and Significance

At present, the overall situation of air quality in our country is not optimistic. The problem of air environmental pollution, with fine particulate matter (PM2.5) as the main pollutant, is becoming increasingly serious. It not only has a great impact on air quality, but also causes serious harm to human health. With the support of big data technology, we can obtain massive air quality data, including concentration of various pollutants, meteorological data, traffic flow, etc. These data provide abundant information sources for the prediction of air quality index and scientific basis for the formulation of environmental policies and the implementation of environmental protection measures. As a classical time series analysis method, ARIMA model has a wide range of applications in the field of air quality prediction. LSTM model, as a special cyclic neural network (RNN), is especially suitable for dealing with long-term dependence problems in sequence data. Therefore, it also has unique advantages in air quality prediction. By comparing ARIMA and LSTM models with rolling prediction, we can predict the change trend of air quality index more accurately. As a tourist resort and an ecological livable city, the air quality of Guilin has always attracted much attention. However, with the acceleration of urbanization and the improvement of industrialization level, Guilin is also facing the problem of air quality decline. Therefore, taking Guilin City as an example, the study on air quality index prediction will not only help to improve the air quality monitoring and early warning capabilities of Guilin City, but also provide reference and reference for air quality prediction of other cities. To sum up, the comparative study of air quality index prediction based on ARIMA and LSTM models has important theoretical significance and practical application value. Taking Guilin as an example, the research background not only reflects the importance and urgency of air quality prediction, but also shows the application prospect of big data and artificial intelligence technology in the field of environmental management.

1.2 Research Status at Home and Abroad

Driven by big data and artificial intelligence, the prediction of air quality index (AQI) has become a research hotspot in the field of environmental science and computational intelligence. ARIMA and LSTM are two commonly used prediction models. They have their own advantages in processing time series data. The better model is determined by rolling prediction. Fang Xiaoping and others took the air quality index (AQI) of 100 cities in the Yangtze River Economic Belt as the research object. Aiming at the characteristics of air quality related data, they established the PSOGSA-LSTM combined prediction model to test the prediction accuracy of the model in three aspects, and compared the prediction results with the traditional LSTM model. Finally, they applied it to the air quality index prediction of 100 cities in the Yangtze River Economic Belt in the next seven days[1]. Based on the monitoring data of air pollutants in Taiyuan City from 2014 to 2017, Zheng Yangyang et al. first used the correlation analysis function in python to analyze the correlation between pollutants and AQI index, and then established the long-term and short-term memory circulation neural network (LSTM) model based on the deep learning library Keras (a

high-level neural network API). The air quality index (AQI) of Taiyuan city is predicted by simulation. The experimental results show that the root mean square error of the model is 4.875, which has the advantages of high prediction accuracy and wide range. It provides a scientific and reasonable theoretical basis and a new prediction method for the prevention and control of air pollution[1]. In addition, the research of rolling prediction model in the field of air quality index (AQI) prediction is currently in a positive development stage. Ling Jin adopts rolling prediction method. In each rolling prediction cycle, the current meteorological data are collected first, and the prediction model adopted is retrained. Then the prediction result is obtained by error superposition correction method.[2] Sun Jing et al. took an index published by the National Bureau of Statistics as the prediction target, and based on the variable screening mechanism, scrolled through the network keyword library to screen explanatory variables, and established a prediction model through various machine learning methods. [3] The purpose of the rolling prediction in this study is also to explore methods to solve such problems as "the complexity and variability of big data on the Internet affect the stability of the model."[4]Ricardo Navares et al. studied the method to predict the concentration of pollutants and pollen, and described that the LSTM neural network model has different topological structures. The prediction results of the LSTM neural network model are more accurate than those of the traditional model.[12]

2 THEORETICAL BASIS

2.1 Principle of ARIMA Model

The ARIMA model is a commonly used time series analysis method, which is used to forecast and model the time series data. It combines autoregressive (AR) model, difference (I) model and moving average (MA) model.[5] The following describes these three parts and their formulas:
The formula for the ARIMA model can be expressed as:

$$\boldsymbol{Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t} \#(1)$$

In this formula (1): is the time series data we are considering. To are the parameters of the AR model, which are used to describe the relationship between the current value and the past P point-in-time values. To are the parameters of the MA model, which are used to describe the relationship between the current value and the errors at the past Q points in time. Is the error term at point in time t. C is a constant term.$Y_t \varphi_1 \varphi_p \theta_1 \theta_q \varepsilon_t$
The idea of ARIMA model is that the time-varying data series are regarded as random series and the change trend is described by mathematical model. The ability to capture the trend and periodicity of time series data to some extent and make predictions based on past observations to provide estimates of future points in time. By adjusting the model parameters, different types of time series data can be modeled and predicted.[6]

2.2 Principles of LSTM Model

LSTM avoids the long-term dependency problem through deliberate design. It is a variant of RNN, which is specially designed to solve the long-sequence dependency problem.[7] It can model long-term dependencies in serial data. LSTM controls the flow of information through the gating mechanism of input gate, forgetting gate and output gate, and simulates the switching function to regulate the input, retention and output of data. LSTM is a PNN model, which is an improvement on Simple RNN. Its structure diagram is shown in the figure, and its principle is similar to Simple RNN[8], the state vector h is updated whenever an input x is read.

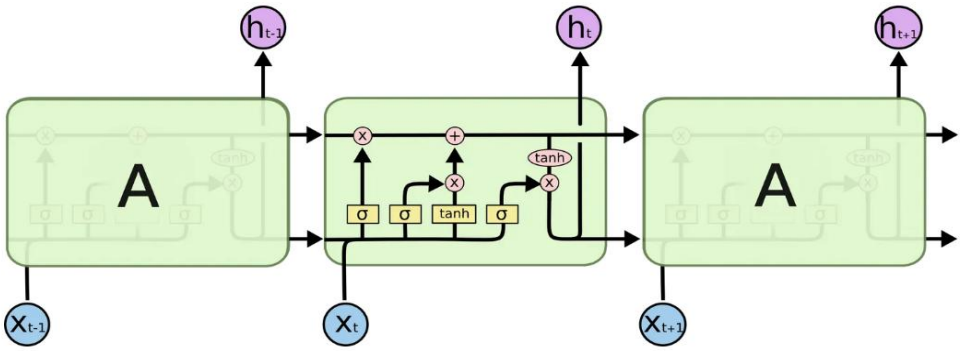

Figure 1 Schematic Diagram of LSTM Network Structure

The calculation process is shown in equations (2) to (8), where F, I and O are forgetting gate, input gate and output gate respectively; C represents short-term memory and h represents long-term memory. Is the activation function; W is the transformation weight matrix (,) from the unit vector to the gate vector; X as the current input; B is the vector feature () obtained by each gate of the input layer. It's a cellular state.$\sigma W_f W_i$、W_o、$W_c b_f$、b_i、b_o、$b_c c_t$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \#(2)$$

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \#(3)$$

$$Sigmoid = 1/e^{-1} \#(4)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \#(5)$$

$$\widetilde{c_t} = \tanh(W_c[h_{t-1}, x_t] + b_c) \#(6)$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \#(7)$$

$$h_t = o_t \tanh(c_t) \#(8)$$

2.3 Principle of Rolling Forecast Model

The principle of Rolling Forecasting is to use historical data to continuously update the forecasting model to adapt to the changing data. Specifically, the rolling prediction is to calculate the state transition matrix P based on the historical data in the previous Ts at each time instant, and predict the vehicle speed state (or other amount to be predicted) for the next second based on the state transition matrix P, where T represents the size of the rolling time window.[11]

In rolling forecasts, the model is retrained each time a new piece of data is added and the value at the next point in time is predicted. In this way, the prediction model can be kept updated and closer to the real data changes, thus improving the accuracy of the prediction. The rolling prediction technique is widely used in time series prediction, especially in scenes requiring real-time prediction, such as traffic flow prediction, energy demand prediction etc.[9]

Rolling prediction also involves some specific model framework principles, such as SCINet, which is a hierarchical down sampling-convolution-interaction TSF framework that can effectively model time series with complex time dynamics. Rolling prediction is a dynamic prediction method based on historical data. It adapts to the changes of data by constantly updating the prediction model, so as to improve the accuracy of prediction.[10]

3 DATA PRESENTATION

3.1 Data Sources

The air quality data in this paper covers the period from 31 December 2013 to 10 May 2024. The air quality index (AQI) is an intuitive indicator of air quality, which is a standardized index to measure air quality. AQI is calculated based on the concentrations of six different pollutants (such as PM2.5, PM10, sulfur dioxide, carbon monoxide, ozone and volatile organic compounds) and converted into a comprehensive index value, which reflects the air quality level and the degree of impact on human health. The data characteristics included are shown in Table 1:

Table 1 Data Characteristics and Dimensions

Data	Characteristic	Dimension
Air pollution data	$PM_{2.5}$ PM_{10} NO_2 SO_2 O_3 CO	6

Generally speaking, a higher AQI value indicates a poorer air quality and a greater impact on health. The relevant information about the air quality index is shown in Table 2. AQI is often used to convey air quality information to the public to help people understand the air quality condition of the day and take corresponding protective measures.

Table 2 Classification of Air Quality Index

AQI value	Air quality level	air quality	represent color
0-50	level 1	excellent	green
51-100	level 2	good	yellow
101-150	level 3	Light pollution	orange
151-200	level 4	Moderate pollution	red
201-300	level 5	Severe pollution	purple

>300	level 6	Serious pollution	maroon

3.2 Data Preprocessing

In this paper, interpolation is used to fill in the missing values in the data set, and the average value of the previous value and the next value in the missing value in the data frame is taken to replace the original missing value. In order to make the prediction more accurate, we extract features and target variables. The feature and target variables are normalized and their values are scaled between 0 and 1. We used the Min-Max normalization model and the Z-score normalization model in the scaling process.[13] By comparing the average mean square error after their processing, we finally chose the Min-Max function to perform the normalization process. Its principle is to use the maximum and minimum values in the data to scale accordingly. The processing formula (9) of Min-Max Scaler is as follows:

$$X_{scaled} = \frac{X - X_{min}}{X_{max-X_{min}}} \#(9)$$

Where x is the original number, the minimum value in the original data, the maximum value in the original data, and the scaled data.$X_{min}X_{max}X_{scaled}$

4 EXPERIMENTAL ANALYSIS

4.1 AQI Forecast of ARIMA Model

4.1.1 Model identification

We can evaluate the autocorrelation and partial autocorrelation of the time series data, so as to preliminarily determine the values of the parameters P and Q in the ARIMA model. Then, we try different combinations of P and Q through grid search method, and use information criteria (such as AIC, BIC, etc.) to select the optimal model. In this paper, the ARIMA (4,1,4) model is determined to be the best fit model after grid search and information criterion evaluation.

4.1.2 Model validation

As shown in Figure 2, the residuals of the ARIMA model show characteristics close to normal distribution, but there is still a slight deviation. This usually means that the model has fitted the data relatively well, although further optimization of the model or the use of different types of models may improve the accuracy of the prediction.

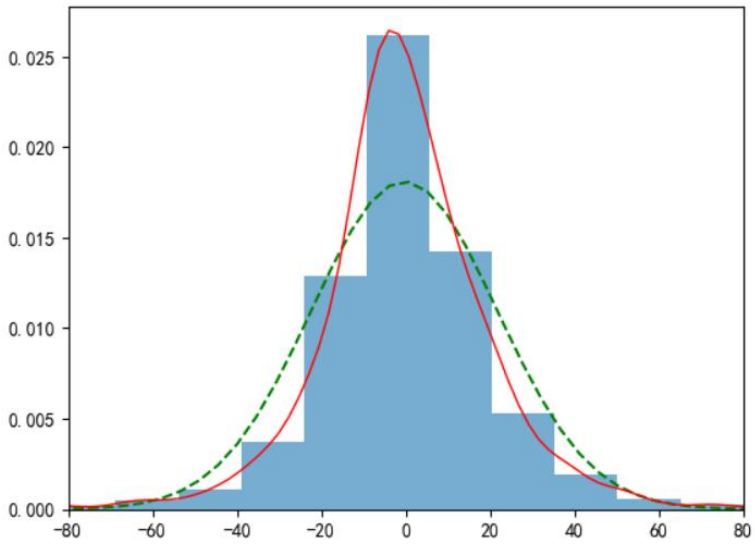

Figure 2 Residual Density Diagram

As can be seen from the QQ plot, the red straight line in the plot represents the expected trend of the theoretical distribution, i.e. if the data completely meets the standard normal distribution, the data points should be roughly distributed along this straight line. The blue dots indicate the correspondence between the actual observations and the theoretical normal quantiles.

Regarding the residual diagnosis of the model, the blue data points are relatively concentrated near the straight line, indicating that the central part of the data is relatively in line with the normal distribution. At both ends of the theoretical quantile (especially at extremes), the data points deviate from the theoretical straight line. In general, the residuals of the ARIMA model are more in line with the normal distribution in the central part, but there is a certain degree of deviation at both ends.

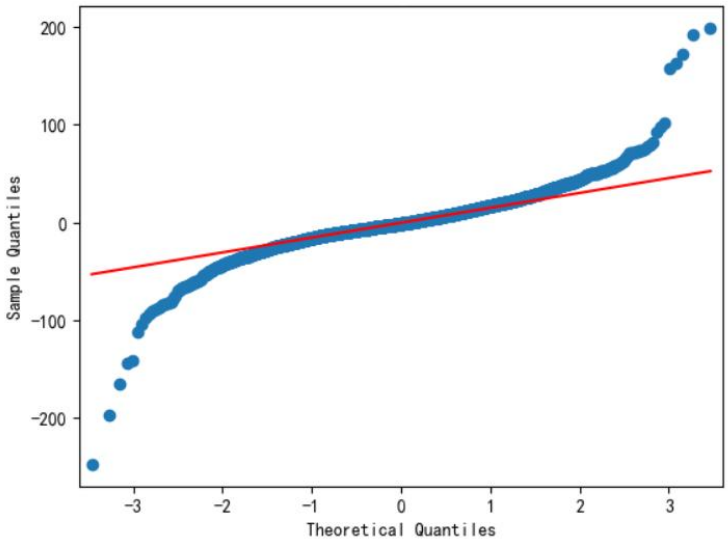

Figure 3 QQ Chart

4.1.3 Model forecast

Comparing the fluctuation between the observed value and the fitted value, as shown in Figure 4, the actual observed curve in blue fluctuates greatly, while the fitted curve in red is relatively smooth. This indicates that the ARIMA model may have a certain smoothing effect when dealing with high-frequency fluctuations, resulting in some extreme values and peaks not being fully captured. The response speed of the model and the red fitting curve delay the response of the blue actual observation curve at some points, especially when the observation value has a big jump or drop. This may indicate that the model has limited adaptability to rapidly changing environmental conditions.

Although there are some local errors and delays in capturing the overall trend, the red fitting curve can generally follow the trend of the blue observation curve. This shows that ARIMA model is effective in capturing the overall trend change of AQI.

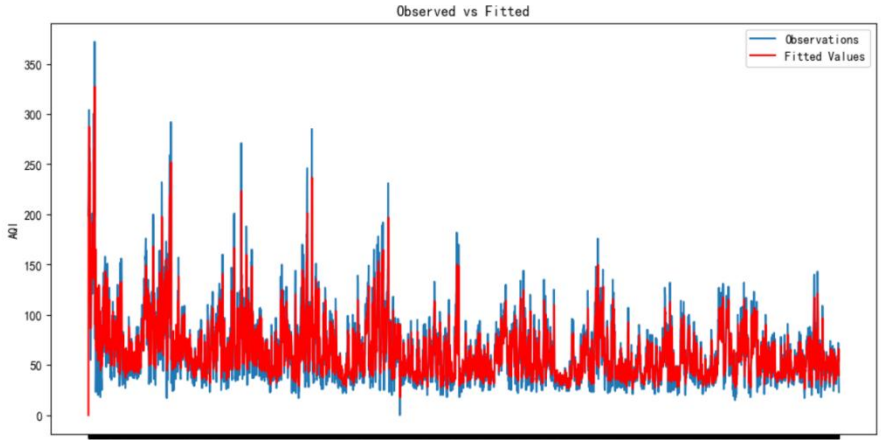

Figure 4 Simulation Forecast Diagram

As shown in Figure 6, it can be concluded from the autocorrelation function (ACF) and partial autocorrelation function (PACF) graphs of the prediction model that the time series data have no significant correlation when the lag order is high. In this case, the ARIMA model does not need autoregressive (AR) part or moving average (MA) again.

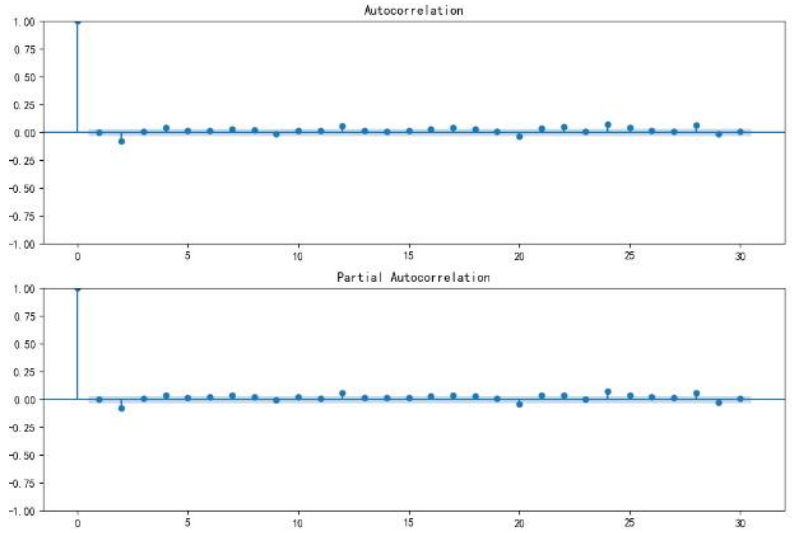

Figure 5 Graph of Autocorrelation Function (ACF) and Partial Autocorrelation Function

Next, the ARIMA (4,1,4) model is used to predict the AQI value of Guilin City for 7 days from May 4, 2024 to May 10, 2024.

As shown in Table 3, comparing the predicted value of AQI with the actual value of AQI, it is found that the absolute error value of the prediction by using ARIMA (4,1,4) model is within (0-10), the relative prediction error is < 25%, and the average relative error is 13.4%. From this, we can draw a conclusion that the prediction accuracy of our model is ideal.

Table 3 ARIMA (4,1,4) Model Predicted Values

date	AQI actual	AQI forecast	error value	Relative Forecast Error/%
2024/5/4	25	30.17502	-5.17502	20.7%
2024/5/5	41	36.33652	4.663481	11.37%
2024/5/6	33	40.10377	-7.103768	21.52%
2024/5/7	52	42.47779	9.522206	18.31%
2024/5/8	52	44.23757	7.762428	14.92%
2024/5/9	49	45.55597	3.444032	7.03%
2024/5/10	46	46.06787	-0.067867	0.15%

4.2 AQI Forecast of LSTM Model

The LSTM neural network model is used to predict the AQI of Guilin City. The data of the first three days are used to predict the air quality of the next day. An LSTM model is constructed, which contains an LSTM layer and a dense connection layer. The mean square error is used as the loss function, and the Adam optimizer is used as the optimizer to fit the model using the training data, which is evaluated on the verification set. Finally, we used Matplotlib to visualize the model's loss values on the training and validation sets in order to evaluate the training of the model. The results are shown in Figure 6 below.

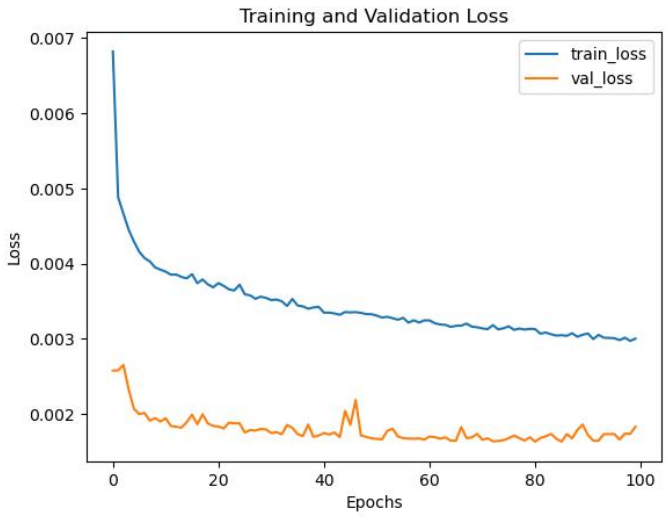

Figure 6 Curve of Loss Change of Model

The graph shows the change of the model's loss values on the training set and the verification set with the training rounds. The abscissa is the training round and the ordinate is the loss value. By comparing the loss curves on the training set and the verification set, it can be evaluated whether the training process of the model is over-fitted or under-fitted. As can be seen from the Figure 7, the training loss (blue line) continues to decline, and the verification loss (orange line) fluctuates slightly and generally shows a downward trend. This indicates that the model is learning from the training data, and there is no obvious over-fitting or under-fitting, indicating that the model can effectively learn the training data.

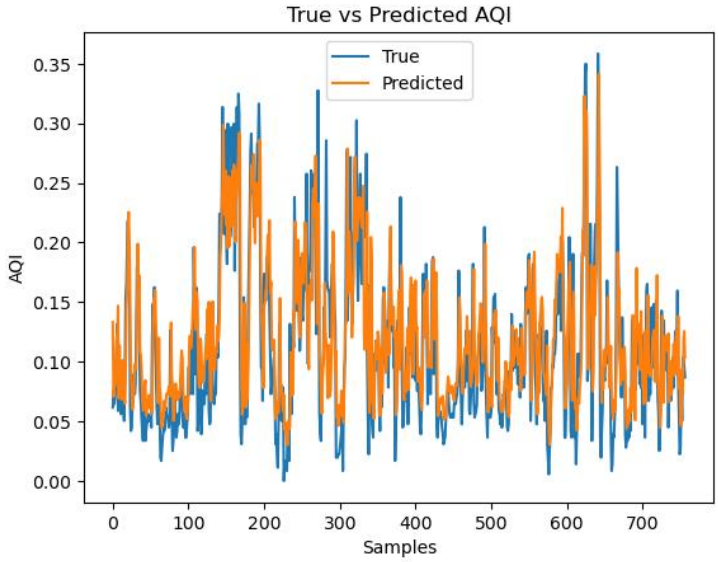

Figure 7 Comparison of Predicted and Actual Values based on LSTM Model

The graph shows the comparison between the predicted results of the model on the test set and the true target value. The abscissa is the index of the sample (i.e. the number of samples) and the ordinate is the AQI value. The graph includes two curves, the blue line segment represents the true AQI value, and the orange line segment represents the AQI value predicted by the model. By observing this graph, the prediction performance of the model can be intuitively evaluated, showing that the predicted value is relatively close to the real value in a number of intervals, especially in the capture of extreme points, which is relatively good, with moderate volatility, and showing good model stability and generalization ability. It shows that our forecast result is accurate. Real-time air quality forecasts can be provided to government departments, environmental protection agencies or the public.

Table 4 Model Evaluation Indicators

evaluating indicator	index value
RMSE	0.03281
MAE	0.02297
MSE	0.00108

As shown in Table 4: RMSE is the standard deviation of the prediction error and represents the average deviation between the predicted value and the actual value. Because the value of RMSE is small, the prediction result of the model is very close to the actual value, and the prediction performance of the model is good. MAE is the average value of the absolute value of the prediction error, which directly reflects the average error between the predicted value and the actual value. MAE is 0.02297, the smaller value of MAE indicates that the prediction error of the model is smaller and the prediction result is more accurate. MSE is the average value of the square of the prediction error and can more sensitively reflect the deviation between the predicted value and the actual value. The value of MSE is 0.00108. The smaller MSE value indicates that the deviation between the predicted result and the actual value of the model is small and the prediction performance is good.

4.3 Rolling Forecast Model Training

The training process of rolling prediction model is usually divided into three stages: first, historical data cleaning and quality verification are carried out to ensure the integrity of time series data; Subsequently, model training is

carried out based on the initial data set, and a basic prediction model is constructed through parameter optimization; Finally, a rolling iteration step is entered, the predicted values generated by the model are backfilled to the training set, the model parameters are dynamically updated with new data, and the prediction error is continuously monitored by using the verification set. In this process, the adaptability of the model to time series changes is gradually improved through the mechanism of data closed loop and model iteration, and the double optimization of prediction accuracy and stability is finally achieved.

4.3.1 ARIMA model forecast

By looking at fig. 9, it can be seen that the predicted value red line roughly follows the trend of the actual value blue. The predicted curve is close to the true value most of the time. However, the volatility of the forecast value is large, especially at some extreme points. This may indicate that the model is overreacting to some specific changes in the data, or that the model has not been able to capture all the factors that affect changes in AQI. Although the overall tracking ability of the model is good, the large fluctuations and deviations indicate that the model still has room for improvement in stability and accuracy.

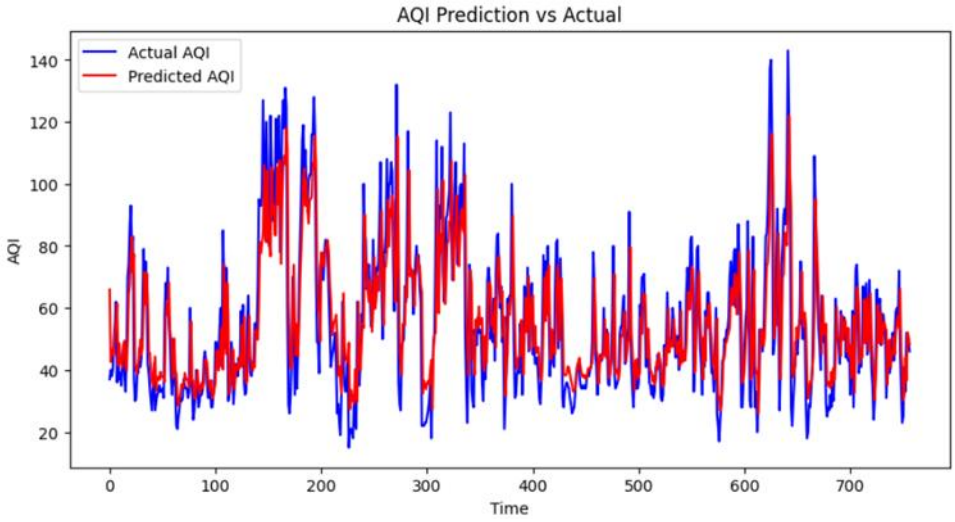

Figure 8 Actual and Predicted AQI Trend of ARIMA Model over Time

4.3.2 LSTM model prediction

By observing fig. 10, it can be seen that the predicted value red line roughly follows the trend of the actual value blue, showing that the model can better capture the change pattern of AQI. This indicates that the prediction model used is effective for trend prediction. This model is very important for environmental management and health early warning system, and can help relevant departments to make timely response and mitigate the impact of adverse air quality.

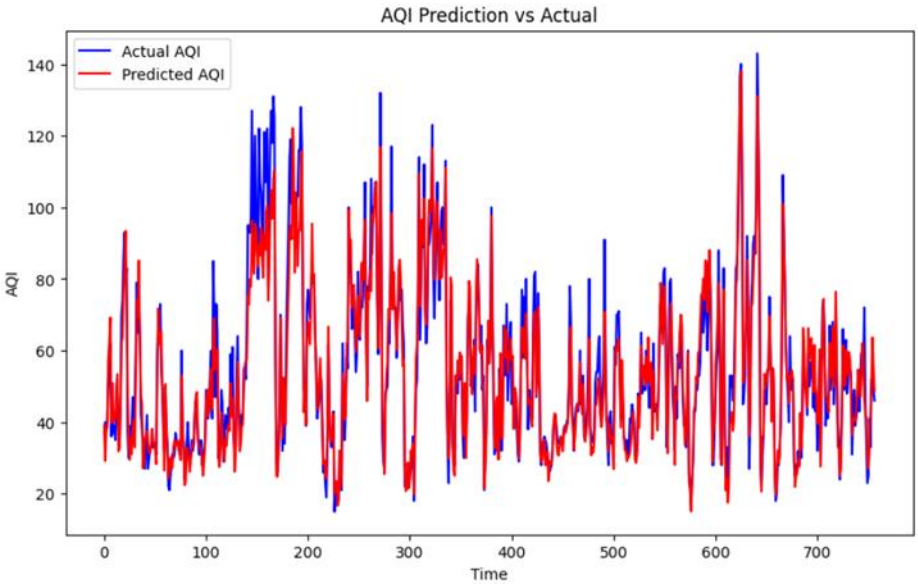

Figure 9 Trend Chart of Actual and Predicted AQI of LSTM Model over Time

4.3.3 Comparative analysis of forecast results

The rolling prediction results under ARIMA model and LSTM model are compared and mainly evaluated by MSE

and RMSE values.

Table 5 Comparison of Model Prediction Errors

Model structure	Test RMSE	Test MSE
ARIMA	14.8031	219.1332
LSTM	7.6096	57.9069

The model is judged according to the mean square error of the model test. The smaller the root mean square error is, the closer the predicted value of the model is to the real value. As can be seen from Table 5, LSTMThe root mean square error of the model is less than that of ARIMA model, indicating that the prediction of AQI index by LSTM model is accurate. Therefore, we choose LSTM model to forecast the future air quality in Guilin.

Considering the above indicators, the rolling forecast AQI under LSTM shows reasonable performance in general, and can effectively predict the results of air quality index of Guilin in the next ten days. As shown in Table 6, it can be concluded that the air quality is in excellent condition, which is very suitable for outdoor activities.

Table 6 Results of Model Forecast AQI

Time	2024.5.11	2024.5.12	2024.5.13	2024.5.14	2024.5.15
Predicted value	40.1942	31.6391	40.7786	46.9876	46.6520
Time	2024.5.15	2024.5.16	2024.5.17	2014.5.18	2024.5.19
Predicted value	45.4955	44.9733	45.0966	45.2633	45.3021

5 CONCLUSION

This paper aims to use ARIMA model, LSTM model and rolling forecast method to forecast the air quality of Guilin city based on big data in order to improve the accuracy and real-time of the forecast. By collecting the air quality data of the mayor's time in Guilin and combining the advantages of the three models, we successfully constructed a model that can accurately predict the air quality change in Guilin. The ARIMA model is used to capture the trend and seasonality of the data, while the LSTM model can better deal with the complex sequence relationship, while the rolling prediction method compares the prediction effects of the two, and updates the real-time parameters to adapt to the changes of the data. Specifically, the ARIMA model plays an important role in predicting the long-term trend and seasonality of the air quality index, while the LSTM model is more suitable for capturing complex dynamic relationships in the data, thus improving the accuracy of the prediction. The rolling prediction model ensures that the model parameters are updated in time to adapt to the changes of new data, thus keeping the real-time prediction.

Through the comprehensive application of these three models, this paper obtains the air quality index prediction results of Guilin with high accuracy and stability. This provides an important reference for air quality management in Guilin and a useful methodology for air quality prediction in other similar cities. In the future, the model parameters and algorithms can be further optimized to improve the prediction accuracy and real-time performance in order to better meet the needs of urban air quality management.

There are still some deficiencies in this paper, among which there may be a certain degree of subjectivity in model selection and parameter setting. The model selection and parameter design should adopt an objective method integrating various factors. In the future, more rigorous model selection and parameter optimization can further improve the prediction performance. On the prediction performance of the model, the mechanism and explanation behind the model may not be explored deeply enough. In the future, the internal mechanism of the model can be further studied in order to better understand the changing laws and influencing factors of air quality.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Fang Xiaoping, Chen Xiuluan, Chu Qi, et al. prediction of air quality index in Yangtze river economic zone based on PSOGSA-LSTM model. Mathematical statistics and management, 2023, 42(01).
[2] Zheng Yangyang, Bai Yanping, Hou Yuchao. Application of LSTM model based on Keras in prediction of air quality index. Mathematical practice and understanding, 2019, 49(07): 138-143.
[3] Ling Jin. Short-term wind speed prediction method based on rolling prediction and its error analysis. Hefei University of Technology, 2017.
[4] Sun Jing, Zhu Jianlin, Li Wanlan, et al. Research on rolling forecast of consumer confidence index based on internet data. Journal of Xi 'an Jiaotong University (Social Science Edition), 2021, 41(06).
[5] Ma Siyuan, Jiao Jiahui, Ren Shengqi, et al. Missing Value Filling of Urban Multiple Air Quality Data Based on Attention Mechanism. Computer Engineering and Science, 2023, 45(08): 1354-1364.
[6] Guo Kuo. Applicability Analysis of ARIMA Model in Precipitation Forecast in Fuxin Area. Heilongjiang Water Conservancy Science and Technology, 2024, 52(04): 37-39+79.
[7] Gu Keming. Research on air quality prediction based on spatio-temporal depth neural network. Zhejiang University of Technology, 2024.
[8] YU Hui Qing, HUANG Peng Cheng, ZHAO Zhen Yu, et al. Standard cell delay prediction method based on RNN. Journal of Zhengzhou University (Science Edition), 2024, 1-7.
[9] Tian Xuwei, Yang Kai, Yin Tong, et al. Prediction of port ambient air quality index based on SSA-Bi-LSTM. Transportation Energy Conservation and Environmental Protection, 2023, 19(05): 32-36+41.
[10] Zhang Tingzhong, Zhang Qinghui, Xing Qiang, et al. Short-term wind speed rolling prediction method based on combination of LMD and GA-BP neural network. Shandong Power Technology, 2019, 46(11): 13-20.
[11] Atakan J, Ayse B O. Forecasting air pollutant indicator levels with geographic model 3 days in advance using neural networks. Expert Systems with Applications, 2010, 37(12): 7986-7992.
[12] A Monterio, M Vieira, C Gama, et al. Miranda. 2017. Towards an improve air quality index. Air Quality, Atmosphere And Health, 10(4): 447-445.
[13] Navares R, Aznarte J L. Predicting air quality with deep learning LSTM: Towards comprehensive models. Ecological Informatics, 2019, 55: 101-119.