Air Quality Prediction Model Based on BP Neural Network and LSTM

Yishuang Li, Zhimin Lin *, Jiayuan Jin, Deshao Kong

School of Mechanical Engineering, Lanzhou Jiaotong University, Lanzhou, China, 730070

* Corresponding Author Email: linzhimin@mail.lzjtu.cn

Abstract. This article uses monthly observation data from national level ground meteorological observation stations in Beijing from 2014 to 2022, and uses LSTM model and ARIMA time series model to predict the data of air quality index factor indicators. A BP neural network model is used to construct a prediction model for Beijing's air quality index. By comparing the predicted data of Air Quality Index (AQI) with the actual data, the accuracy of the model's predicted AQI in Beijing is over 95%, indicating a relatively accurate prediction accuracy. It can be used for predicting the air quality index in Beijing. Provide technical support for air quality and environmental governance in Beijing.

Keywords: BP Neural Network, Air Quality Index Prediction Model, LSTM Model, ARIMA Time Series Model.

1. Introduction

Air Index (AQI) is an important index to quantitatively describe air conditions. With the process of my country's modernization and industrialization, air pollution has intensified and air has declined. In response to my country's air problems, the State Council issued the "Air Improvement Plan", which reflects my country's attitude towards air problems. Therefore, how to scientifically predict the air index of various regions has very important significance.

Aiming at the accuracy and regional problems of forecasting air index models. Zhang et al. applied machine learning based on random forest algorithm to air index prediction, realizing pollution refinement to town-level units, making the measurement results closer to the real value [1]. Considering the complexity of the model, Haq proposed the SMOTEDNN model, which effectively improved the accuracy of AQI prediction [2]. Udristioiu et al. incorporated the concentration of polluted particulate matter, considered the direct factors affecting AQI, integrated IVS, machine learning, and regression prediction-based decision tree machine learning recognition (NARMAX), and the error margin was greatly reduced [3]. Sun et al. established a comprehensive model of CEEMDAN-ARMA-LSTM to improve the utilization rate of collected data and significantly improve the prediction accuracy of AQI [4]. For the limited prediction performance, Ji et al. used CEEMDAN technology to propose a two-scale integrated framework prediction, which predicts high-frequency and low-frequency components separately, so that the prediction has a degree of discrimination [5]. Li et al. proposed a mixed model of TVFEMD = SE, GWO, DLH, which made the prediction accurate to every hour and improved the dynamics of the prediction [6]. Duong et al. designed a TLPW element based on sensor data and stacking generalization, adding more influencing factors to make the prediction results more regionally accurate [7].

Although the current models for predicting AQI have integrated and deepened algorithms, most of them do not consider regional factors and meteorological factors, such as average monthly precipitation, or do not consider regional differences and AQI prediction fluctuations brought about by seasonal changes in each region. In view of the influence of climate and pollutant concentration on AQI, this model proposes an air prediction model based on BP neural network. This model extracts Beijing's pollutant concentration and meteorological data through LSTM and ARIMA time series prediction, and predicts Beijing's AQI. Compared with previous models, this model adds more and more complete pollutant factors and meteorological factors, making the prediction more accurate. It can also predict the AQI of the region according to a specific region, and because the added meteorological factor data can be accurate to the month, the AQI prediction of the model can consider regional seasonal changes, greatly reducing error and generalization.

2. Methodology

2.1. Description ofthe Dataset

In the research process of this article, two indicators, namely the degree of air pollution in Bei Jing and the meteorological changes in each city, were selected as the core indicators for predicting and evaluating the air quality index of Beijing, China.

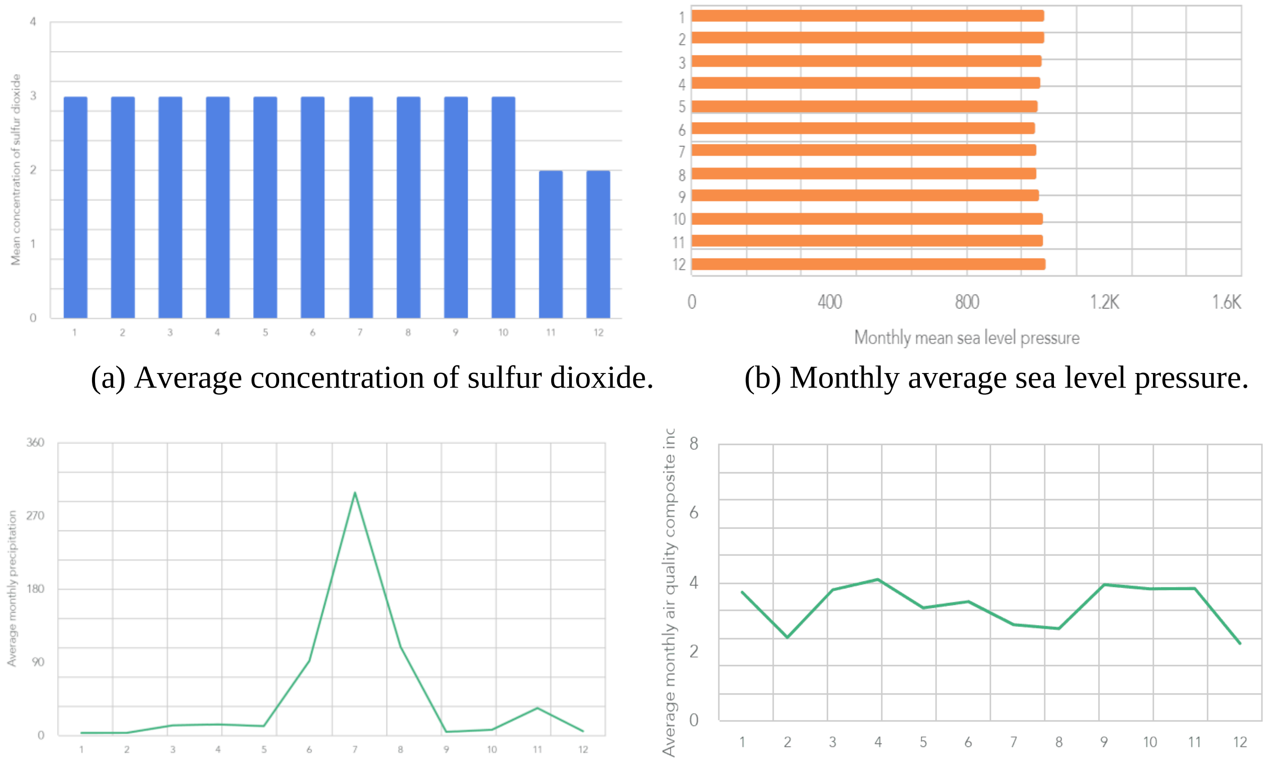
The degree of air pollution in each city is reflected by the AQI value, which is calculated based on the concentrations of several major harmful substances such as PM2.5, PM10, sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and ozone (O_3). The concentration of harmful substances can directly reflect the quality of the air, and specific changes in harmful substances can lead to direct changes in AQI.

The meteorological indicators of a city are reflected by its monthly meteorological data, and meteorological changes can affect processes such as pollution volatilization, diffusion, and deposition.

This article uses data published by the National Bureau of Statistics and RESSET website as research data. The data used is monthly data from national level ground meteorological observation stations in Beijing from 2014 to 2022. The research data includes concentrations of sulfur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3), PM2.5, as well as multiple indicators such as precipitation and wind speed in Beijing.

The data comes from the National Bureau of Statistics and RESSET website. The data is accurate, rigorous, and authoritative, which can provide data support for the research in this article. Data source website: https://www.stats.gov.cn/; https://www.resset.com/index/home/.

Selecting some of the data for visualization processing yields the result shown in Figure 1.



(a) Average concentration of sulfur dioxide. (b) Monthly average sea level pressure.

(c)Monthly average precipitation. (d) Monthly Average Air Quality Composite Index.

Figure 1. Data visualization display.

2.2. Problem Formulation

Using mathematical and statistical methods to analyze the monthly changes in Beijing's air quality index, using ARIMA time series models to predict changes in meteorological data, LSTM models to predict changes in harmful substance concentrations, and then using BP neural networks to construct an air quality index prediction model and test the prediction results.

2.3. Algorithm Introduction

2.3.1. bp neural network model

The study used BP neural network to construct an air quality index (AQI) prediction model. The BP neural network model was proposed by Rumelhart and McClelland in 1986, and is a classic multi-layer feedforward network with the core algorithm being backpropagation. This type of network typically includes an input layer, one or more hidden layers, and an output layer. It calculates the output through forward propagation and continuously adjusts the weights and thresholds of the network through backpropagation to minimize the sum of squared errors of the network [8].

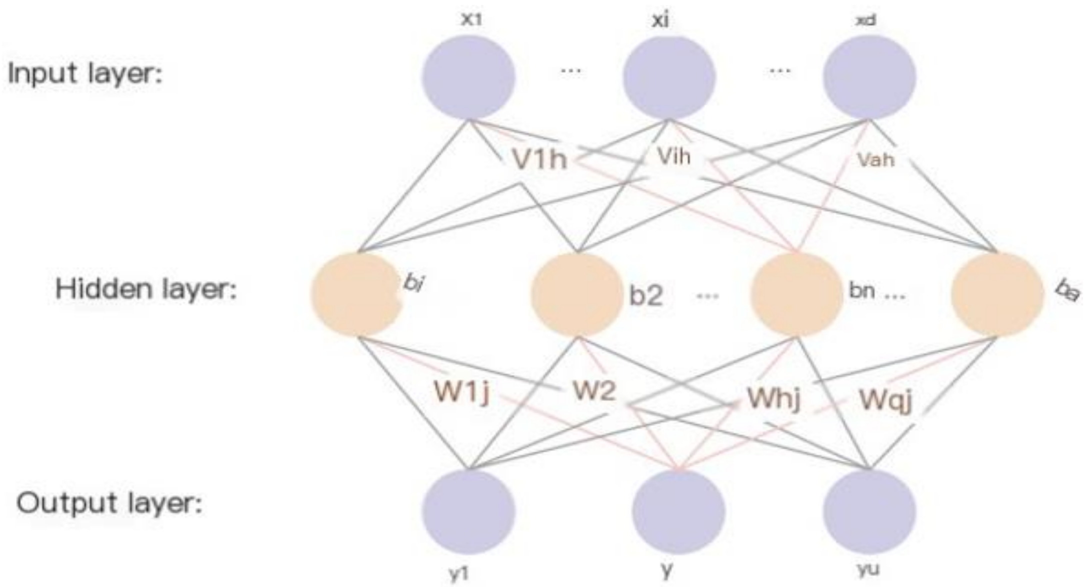


Figure 2. BP neural network structure diagram.

As shown in Figure 2, the number of neurons in the input layer, hidden layer, and output layer are d q, l; X_i is the initial input of the i-th neuron, vih is the weight of the i-th neuron in the input layer and the h-th neuron in the hidden layer, α h is the input of the h-th neuron in the hidden layer, γ h is the threshold of the h-th neuron in the hidden layer, bh is the output of the h-th neuron in the hidden layer, whj is the weight of the h-th neuron in the hidden layer and the j-th neuron in the output layer, β j is the input of the j-th neuron in the output layer, θ j is the threshold of the j-th neuron in the output layer, and yj is the output of the j-th neuron in the output layer.

2.3.2. Long Short Term Memory (LSTM) Model

When traditional RNNs handle long-term dependency problems, there are situations where gradients disappear or explode, making it difficult to effectively model long-term temporal dependencies. To address this issue, a more powerful recurrent neural network has been proposed, namely the Long Short Term Memory (LSTM) neural network model. The LSTM model can selectively remember, update, and output information through gating mechanisms, including forget gates, input gates, and output gates, in order to better capture long-term dependencies. The LSTM model maintains a memory unit within the unit to store and transfer information, demonstrating stronger memory and modeling capabilities when processing sequential data. The schematic structure of the LSTM model is shown in Figure 3 [9].

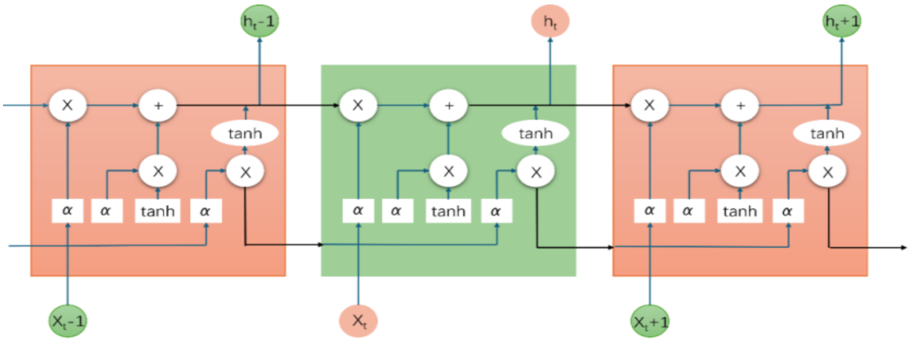


Figure 3. Schematic diagram of LSTM structure.

2.3.3. ARIMA time series model

The ARIMA (Autoregressive Moving Average) time series model is a classic time series analysis and prediction model first proposed by American statistician Box and British statistician Jenkins in 1970. A time series is a collection of data points or observations arranged in chronological order, and the ARIMA time series model is suitable for fitting data that exhibits non stationarity. Its basic idea is to extract patterns hidden behind the data for prediction by combining autoregression and moving average terms, and appropriately performing differential processing on the data.

The ARIMA time series model is divided into two parts: autoregression (AR) part: The AR part describes the relationship between current values and historical values. In the ARIMA time series model, p represents the number of autoregressive terms, which means the model takes into account the impact of p lagged observations on the current value. Moving Average (MA) section: This section focuses on the impact of prediction errors on future values. In ARIMA time series models, 'q' represents the number of moving average terms, which means the model considers the impact of q past prediction errors on the current value.

2.4. Model establishment

Firstly, the LSTM model and ARIMA time series model are respectively used to predict the concentration of harmful substances and meteorological data. Then, by training the BP neural network model, a model that can predict AQI is obtained, and the predicted results are compared with the actual AQI data to determine whether the BP neural network model can be used for prediction. If the BP neural network model can be used for prediction, the BP neural network model and the predicted meteorological data and average concentration of harmful substances are used to predict the AQI of Beijing. If BP neural network cannot be used for prediction, search for other models with higher fitting degree, retrain and predict.

The flowchart of the model construction process in this article is shown in Figure 4.

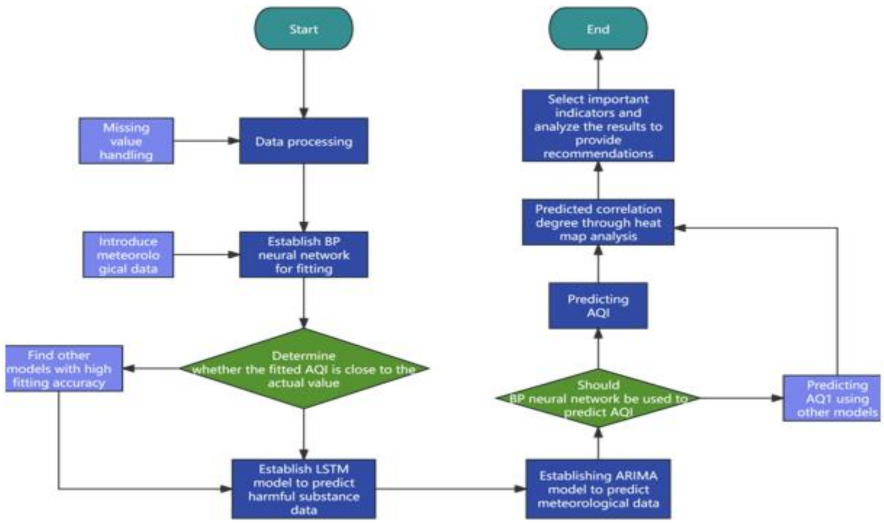


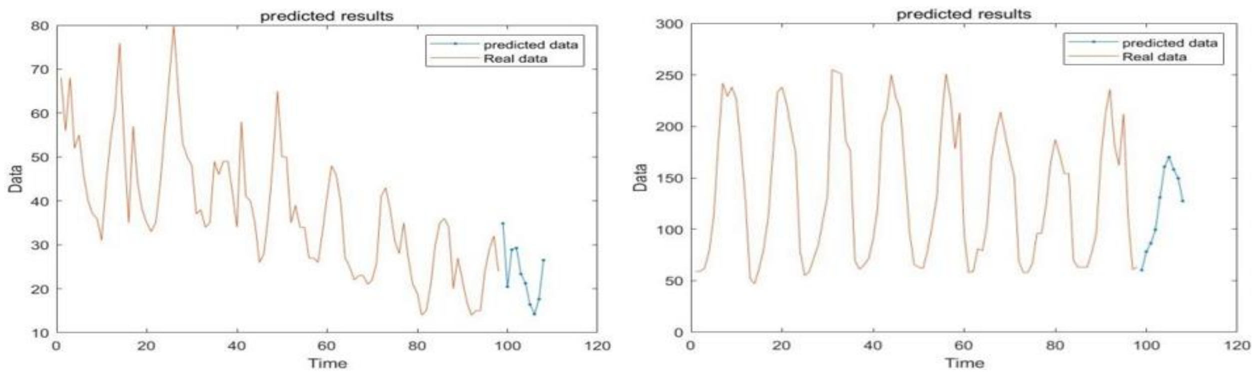
Figure 4. Overall Model Construction Process Diagram.

As shown in the above figure, this article first introduces the monthly meteorological data, harmful substance concentration data, and air quality index of Beijing from 2014 to 2022, and uses the mean filling method to fill in the missing data. The BP neural network model is used for fitting to determine whether the fitting degree predicted by the BP neural network model is consistent with the prediction of the air quality index. If BP neural network models can be used, then LSTM models can be established to predict the monthly concentration data of harmful substances for the next year, and ARIMA time series models can be used to predict the monthly meteorological data for the next year. Finally, the predicted harmful substance concentration data and meteorological data are introduced, and the BP neural network model is used to predict AQI data. Based on the correlation between each indicator and AQI, as well as national environmental governance policies, corresponding environmental governance strategies are proposed by region.

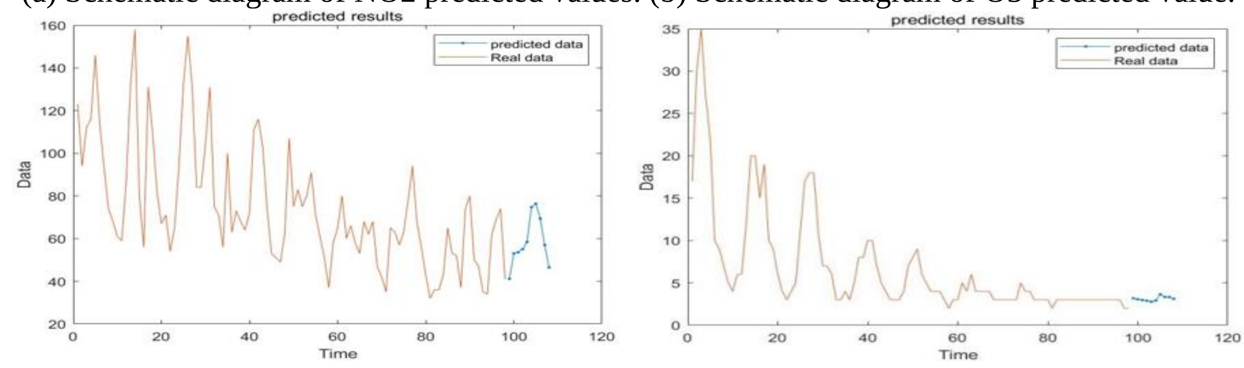
3. Result and discussion

3.1. LSTM model prediction results

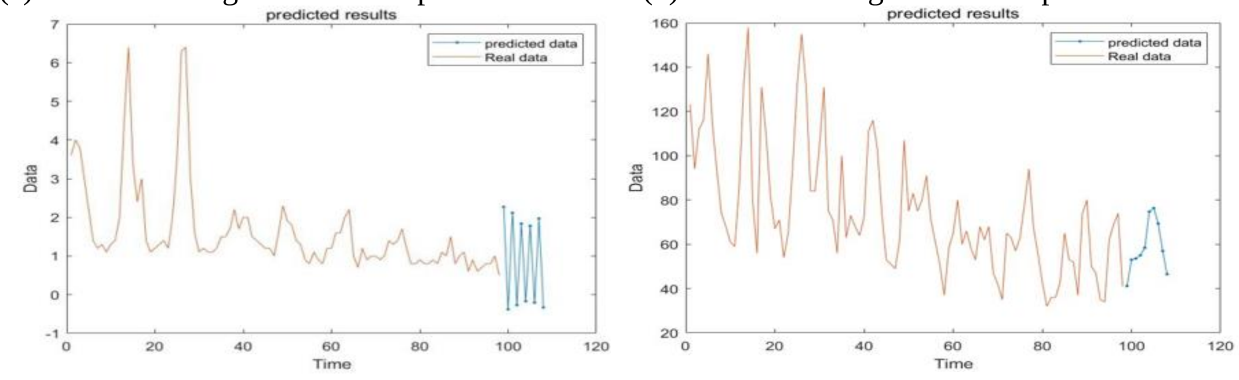
This article uses the LSTM model [1] to predict the concentration of pollutants in Beijing, as shown in Figure 5.



(a) Schematic diagram of NO2 predicted values. (b) Schematic diagram of O3 predicted value.



(c) Schematic diagram of PM10 predicted values. (d) Schematic diagram of SO2 predicted values.



(e) Schematic diagram of CO prediction value. (f) Schematic diagram of PM2.5 predicted values.

Figure 5. Prediction of Six Hazardous Substances Based on LSTM

Figure 5 shows the predicted monthly average concentrations (in ug/m3) of six harmful substances, as shown in Table 1.

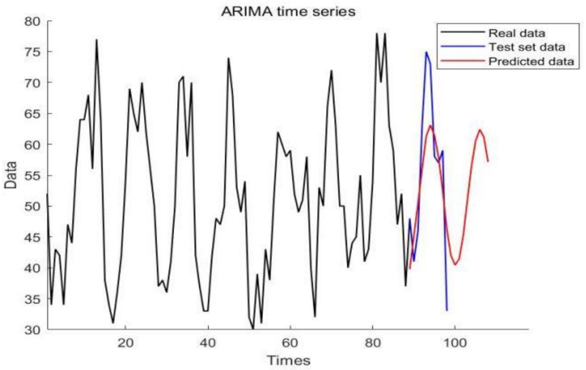
Table 1. Predicted Monthly Average Concentrations of Six Hazardous Substances.

Date	O3	CO	NO2
2023/1/31	54.39	0.50	34.64
2023/2/28	78.95	1.36	20.30
2023/3/31	84.43	0.54	29.67
2023/4/30	104.89	0.93	30.17
2023/5/31	154.28	0.86	22.85
2023/6/30	191.27	0.73	18.94
2023/7/31	190.80	0.99	16.42
2023/8/31	167.97	0.78	16.29
2023/9/30	162.68	1.19	21.21
2023/10/31	97.58	0.78	28.99
Date	PM2.5	PM10	SO2
2023/1/31	34.41	65.11	3.09
2023/2/28	39.24	52.92	3.20
2023/3/31	33.01	52.04	3.20
2023/4/30	36.59	37.39	3.10
2023/5/31	24.10	61.10	3.00
2023/6/30	32.81	98.49	2.92
2023/7/31	53.06	77.14	2.82
2023/8/31	75.79	66.95	2.79
2023/9/30	41.26	54.38	2.93
2023/10/31	17.27	45.66	3.08

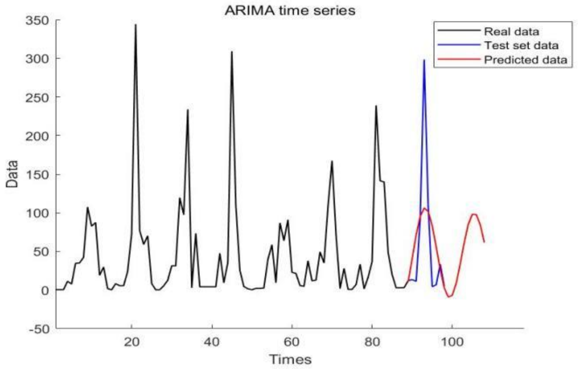
According to China's "Environmental Air Quality Standards", the average concentration limit of O_3 is 160ug/m3, the average concentration of CO should not exceed 4ug/m3, the average concentration of NO2 should not exceed 40ug/m3, the average concentration of PM2.5 should not exceed 35ug/m3, the average concentration limit of PM10 is 70ug/m3, and the average concentration limit of SO_2 is 60ug/m3. According to the table, the average concentration of the vast majority of harmful gases meets the maximum allowable value, but the average concentrations of O_3 from May to September, PM2.5 from February, April, July, August, and September, and PM10 from June and July are relatively high. It is necessary to strengthen environmental protection and propose policy support to achieve the goal of reducing the average concentration of harmful substances.

3.2. ARIMA model prediction results

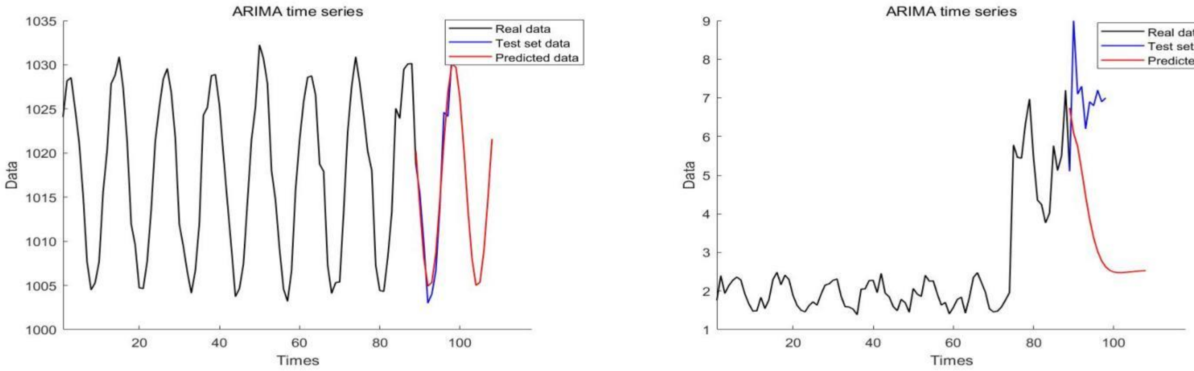
In this study, the initial 80% of the dataset was divided into a training set to build and optimize the model, while the remaining 20% was used as a testing set to evaluate the performance of the model. The results obtained are shown in Figure 6.



(a)Wind speed prediction chart.



(b) Precipitation forecast map.



(c)Air pressure prediction chart. (d) Humidity prediction chart.

Figure 6. ARIMA time series prediction of Beijing meteorological data.

From Figure 6, it can be seen that the Beijing meteorological data prediction curve. The specific prediction results are shown in Table 2.

Table 2 Prediction Results of Four Meteorological Indicators in Beijing

Date	Monthly average relative humidity in Beijing	Monthly average precipitation in Beijing
2023/1/31	42.47	0.00
2023/2/28	40.47	0.00
2023/3/31	41.38	8.87
2023/4/30	44.92	33.94
2023/5/31	50.22	61.68
2023/6/30	55.80	84.96
2023/7/31	60.12	97.96
2023/8/31	62.07	97.57
2023/9/30	61.16	84.16
2023/10/31	57.63	61.39
Date	Monthly average wind speed in Beijing	Monthly average sea level pressure in Beijing
2023/1/31	2.53	1029.97
2023/2/28	2.49	1026.47
2023/3/31	2.47	1020.49
2023/4/30	2.47	1013.63
2023/5/31	2.48	1007.74
2023/6/30	2.49	1004.41
2023/7/31	2.50	1004.53
2023/8/31	2.51	1008.08
2023/9/30	2.52	1014.09
2023/10/31	2.52	1020.95

3.3. Prediction results of BP neural network model

Based on LSTM prediction of six harmful substances and ARIMA prediction of four meteorological data, the results obtained through BP neural network are shown in Figure 7.

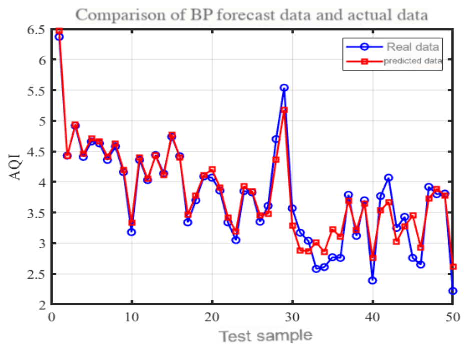


Figure 7. The fitting results of BP neural network on Beijing AQI data.

According to the BP neural network prediction, the average absolute error (MAE) between the predicted and actual data is 0.157, the mean square error (MSE) is 0.047, the root mean square error (RMSE) is 0.217, the correlation coefficient R is 0.97, and the prediction accuracy is 95.08%. Therefore, the BP neural network can be used to predict the AQI of six harmful substances and gas image data.

Using BP neural network prediction, the final AQI prediction results for Beijing are shown in Figure 8 and Table 3.

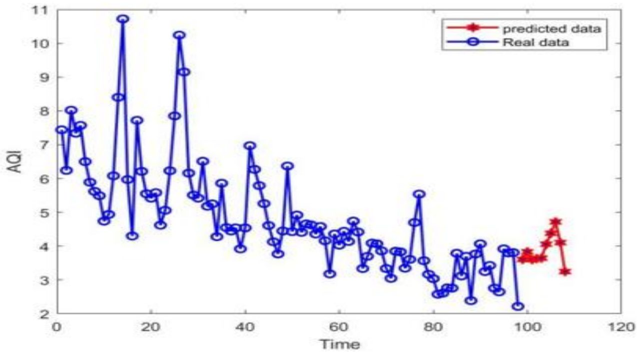


Figure 8. Beijing AQI prediction results.

Table 3. Beijing AQI Predicted Values.

Date	AQI
2023/1/31	3.93
2023/2/28	4.06
2023/3/31	3.89
2023/4/30	4.06
2023/5/31	4.40
2023/6/30	4.92
2023/7/31	5.07

3.4. Result analysis

Based on the data from the Beijing National Meteorological Station from 2014 to 2022, analyze the changes in Beijing's air quality index; Using the pollution concentration data and meteorological data of Beijing from 2014 to 2022, LSTM, ARIMA time series, and BP neural network methods were used to conduct monthly training on the prediction of Beijing's air quality index. A prediction model for Beijing's air quality index was established, and the main results are as follows:

(1) Based on the predicted data, the conclusion can be drawn that there is a significant correlation between air quality and precipitation, relative humidity, and monthly average concentration. But it is relatively dry, with uniform wind speeds throughout the year, and in February, March, April, and May, the average concentration of PM2.5 is higher in February, March, April, July, and September.

(2) Suggestions for the governance of Beijing are as follows: due to the high concentration, it is necessary to strengthen the management of dust on urban construction sites and roads, and implement dust reduction measures such as covering and spraying; Strengthen the greening construction of cities and surrounding areas, increase vegetation coverage, trees and grasslands can effectively absorb PM2.5 particles, and reduce CO_2 concentration through photosynthesis; Strengthen the supervision of industrial emissions, especially the strict control of PM2.5 generated from coal combustion. Promote the use of clean production technologies and energy-saving and emission reduction measures.

(3) The ARIMA time series model can accurately capture the data trends and seasonal changes of time series through differential and lag operations, making meteorological data prediction results highly accurate and reliable, and subject to seasonal fluctuations.

(4) Compared to RNN, the LSTM long short-term memory network model has made significant improvements in solving the gradient vanishing problem during backpropagation. It can effectively

avoid such problems, simulate and train longer sequences, make more accurate predictions, and through its three gate structure, it is very flexible in processing information.

(5) The overall error of the model is small, and there are many factors to consider, making the research and promotion significant; The air quality index predicted by the model has an accuracy of over 95% in Beijing, with an average absolute error (MAE) of less than 0.16, a mean square error (MSE) of less than 0.05, and a root mean square error (RMSE) of less than 0.22 between the predicted and actual data

(6) Some predictions in the model can only make short-term predictions, such as the ARIMA time series model, which can only consider the lag and difference of data and cannot pay attention to other influencing factors, resulting in the ARIMA time series model only being able to make short-term predictions and having weak long-term prediction capabilities.

4. Conclusion

This article uses BP neural network model, LSTM model, and ARIMA time series model to predict the impact indicators of air quality index: pollutant concentration and climate. And by fitting the predicted air quality index (AQI) and comparing it with real data, the prediction accuracy was 95.08%. Therefore, this model can accurately predict the air quality index (AQI) considering pollutant concentration and climate, and provide effective governance suggestions based on the predicted data.

There are also some limitations in this study, such as the ARIMA time series model, which can only consider the lag and difference of data and cannot pay attention to other influencing factors. Therefore, it can only make short-term predictions on meteorological data, resulting in the final Air Quality Index (AQI) being able to make short-term predictions with a relatively short prediction time. In future research, the prediction cycle length can be increased by improving the ARIMA time series model or combining it with other models to propose more comprehensive and accurate models.

References

[1] Zhang yuan Z ,Jia W ,Nina X , et al.Air Pollution Exposure Based on Nighttime LightRemote Sensing and Multi-source Geographic Data in Beijing[J].Chinese GeographicalScience,2023,33(02):320-332.

[2] Haq M A. SMOTEDNN: A novel model for air pollution forecasting and AQI classification [J]. Computers, Materials & Continua, 2022, 71(1).

[3] Udristioiu M T, Mghouchi Y E L, Yildizhan H. Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning [J]. Journal of Cleaner Production, 2023, 421: 138496.

[4] Sun Y, Liu J. Aqi prediction based on ceemdan-arma-lstm[J]. Sustainability, 2022, 14(19): 12182.

[5] Ji C, Zhang C, Hua L, et al. A multi-scale evolutionary deep learning model based on CEEMDAN, improved whale optimization algorithm, and regularized extreme learning machine and LSTM for AQI prediction [J]. Environmental research, 2022, 215: 114228.

[6] Li Y, Peng T, Hua L, et al. Research and application of an evolutionary deep learning model based on improved grey wolf optimization algorithm and DBN-ELM for AQI prediction[J]. Sustainable Cities and Society, 2022, 87: 104209.

[7] Duong D Q, Le Q M, Nguyen-Tai T L, et al. An effective AQI estimation using sensor data and stacking mechanism [M]//New Trends in Intelligent Software Methodologies, Tools and Techniques. IOS Press, 2021: 405-418.

[8] Zhao Z, Xin H, Ren Y, et al. Application and comparison of BP neural network algorithm in MATLAB[C]//2010 International Conference on Measuring Technology and Mechatronics Automation. IEEE, 2010, 1: 590-593.

[9] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures [J]. Neural computation, 2019, 31(7): 1235-1270.

[10] Shumway R H, Stoffer D S, Shumway R H, et al. ARIMA models [J]. Time series analysis and its applications: with R examples, 2017: 75-163.