ISSN 2616-5775 Vol. 6, Issue 10: 113-118, DOI: 10.25236/AJCIS.2023.061017

Prediction of Air Quality in the Beijing-Tianjin-Hebei Region Based on LSTM Model

Na Xu[1,*], Lu Li[1], Hanxiao Dong[1], Feiyang Huang[2]

[1]School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang, 471000, China
[2]College of Mechanical and Electrical Engineering, Henan University of Science and Technology, Luoyang, 471000, China
[*]Corresponding author: xunaydx6@163.com

Abstract: *The air environment plays a vital role in human life and is closely related to the soundness of the ecosystem and the safety of human life, and good air quality is one of the prerequisites for the sustainable development of cities and society. In this paper, the Beijing-Tianjin-Hebei region is selected as the research object to explore the regional air quality characteristics, predict air quality changes, and seek scientific and effective methods and suggestions to improve air quality. In this paper, an air quality prediction model based on the long- and Long Short-Term Memory Networks (LSTM) is established by using the daily average AQI data of six cities in the Beijing-Tianjin-Hebei region for a total of 1,953 days from January 1, 2018, to April 30, 2023, respectively. Finally, the established model was evaluated using several evaluation metrics such as root mean square error (RMSE). The results show that the LSTM-based neural network can predict the AQI more accurately, which provides a scientific and reasonable theoretical basis and prediction method for the environmental protection and related decision-making of governmental departments.*

Keywords: *Air quality, AQI data, Long Short-Term Memory Networks*

1. Introduction

In recent years, the air quality problem has attracted more and more attention, and the pollution of the environment has led to great changes in people's lives, affecting the development of society and the general public to a certain extent, so it is urgent to study the topic of environmental quality. Since the air quality index varies significantly among Chinese cities, and different regions present different characteristics for different reasons, the Beijing-Tianjin-Hebei region is considered as a representative region to be studied and analyzed [1].Beijing-Tianjin-Hebei region as one of the key development areas, the air quality index of this region to carry out relevant research on the environmental management of other cities has a certain reference significance, which is conducive to the realization of the Chinese-style modernization, and promote the harmonious coexistence of man and nature.

There are many methods for air quality prediction, among which the relatively common one is prediction by time series modeling. Peng Sijun [2] and others used the autoregressive moving average model (ARIMA) to predict the short-term daily average concentration of PM2.5 by using the data provided by the environmental monitoring station for the known characteristics of the time series distribution of PM2.5 concentration changes. Youyou et al [3] used a BP neural network optimized by Bayesian regularization (BR) algorithm to predict the air quality condition in Wuhu city, which improved the generalization ability of the network. He ZY et al [4] used the average concentrations of six pollutants to analyze the influence of each pollutant on AQI by constructing a linear quantile regression model. In the process of comprehensively promoting the modernization, the air environment pollution problem cannot be ignored, this paper for the Beijing-Tianjin-Hebei region in Beijing, Tianjin, Langfang, Baoding, Tangshan, Shijiazhuang six representative cities based on the air quality data related to the analysis and prediction. This paper chooses to use the method of artificial neural network to establish the long and short-term memory neural network to predict the air quality index in the Beijing-Tianjin-Hebei region.

ISSN 2616-5775 Vol. 6, Issue 10: 113-118, DOI: 10.25236/AJCIS.2023.061017

2. The basic fundamentals of Long Short-Term Memory Networks (LSTM)

2.1 Data Acquisition and Preprocessing

In this paper, six cities, namely Beijing, Tianjin, Baoding, Langfang, Shijiazhuang and Tangshan, are selected as representative cities to study and analyze air quality related to the Beijing-Tianjin-Hebei region. The air quality data used in this paper are from the China Environmental Monitoring website (https://air.cnemc.cn:18007/), and the daily AQI data of 1953 days from January 1, 2018 to April 30, 2023 are selected as the research object, and the trend of the daily average AQI values of the six cities can be obtained as follows Figure 1.

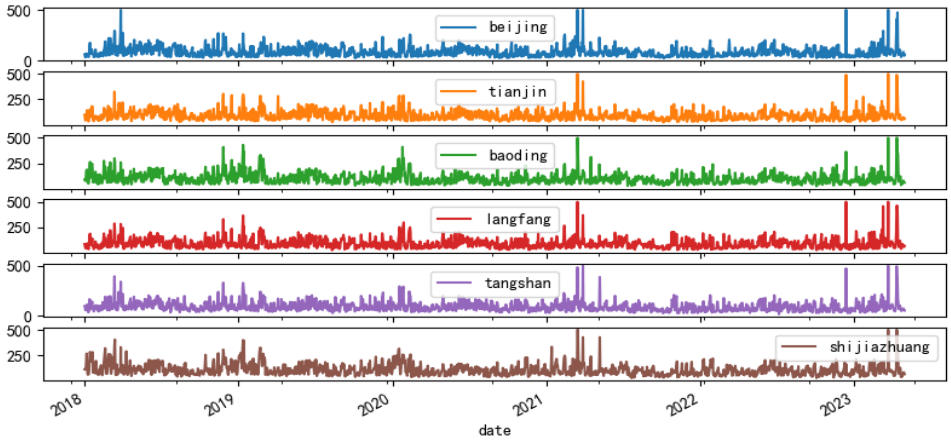

Figure 1: Line graph of daily average AQI values by city

By observation and comparison, the average daily AQI values of Baoding and Shijiazhuang are relatively more volatile, with most of the average daily AQI values of Beijing, Tianjin, Langfang, and Tangshan in the range of less than 150, i.e., less than mildly polluted, while Baoding and Shijiazhuang are relatively more polluted, with more days with average AQI values exceeding 150. The changes in air quality in each city can be visualized in Figure 1, which can lay the foundation for further analysis.

Due to the characteristics of various types of data with different quantities and properties, in order to avoid the situation of slower model training and larger training error due to the large difference in the quantities of the input and output data, this paper adopts the Min-MAX method to standardize the input and output data, and reduces the characteristics of the data to between [0, 1], and the normalized data are the fastest in the search for optimal solutions.

2.2 Principles of LSTM neural network model

LSTM is a special kind of recurrent neural network (RNN), which uses a unique gating mechanism to effectively solve the problem of gradient vanishing during the training of long sequences and the problem of gradient explosion of RNN when the number of iterations increases. It is highly dependent on long data sequences and has strong nonlinear fitting ability. Compared with ordinary RNN, LSTM has better performance in long sequences, and its unit structure is shown in Figure 2.

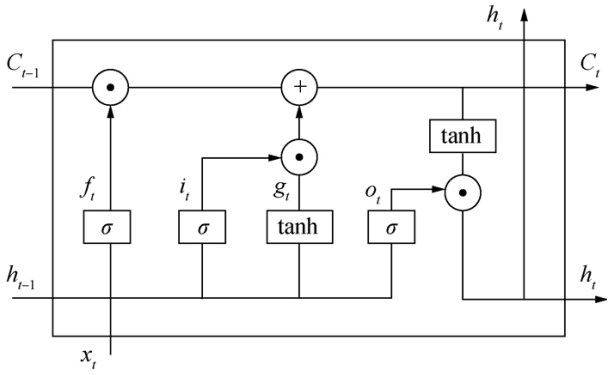

Figure 2: Main structure of LSTM model

ISSN 2616-5775 Vol. 6, Issue 10: 113-118, DOI: 10.25236/AJCIS.2023.061017

In Figure 2, x_t represents the data information at that moment, σ is the sigmoid activation function, and the activation function σ and tanh expressions generally take the following values:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2}$$

C_{t-1} and C_t are the information states of the cellular unit in the previous period and the current moment, and the cellular state is the key to the LSTM neural network. In LSTM, the gate structure has the ability to decide whether the information passes through or not, and the LSTM model realizes the selective deletion or retention of information through three gating mechanisms, i.e., forgetting gate, remembering gate and output gate. Its main formula is as follows [5].

(1)In order to avoid too much memory interfering with the neural network's processing of the current input, we should selectively ignore certain components of the previous unit state in order to determine what information can be passed on, this function corresponds to the forgetting gate.

$$f_t = \sigma(w_1^f \cdot x_t + w_h^f \cdot h_{t-1} + b_f) \tag{3}$$

(2) A memory gate is a control unit that controls whether or not data at a point in time (the current) is merged into a cell state. First, the tanh function is used to extract useful information from the current vector, and then the sigmoid function (to the left of the tanh function layer on the graph) is used to control "how much" of this memory is put into the cell state. The new cell state C_t is saved by first multiplying the cell state C_{t-1} from the previous stage by f_t to forget about the information that we don't need, and then adding the new candidates to the cell state.

$$i_t = \sigma(w_1^i \cdot xt + w_h^i \cdot ht - 1 + b_i) \tag{4}$$

$$\widetilde{C_t} = \tanh(w_1^c \cdot xt + w_h^c \cdot ht - 1 + b_c) \tag{5}$$

$$Ct = ft \cdot Ct - 1 + it \cdot \overline{C}_t \tag{6}$$

(3) The output of the model relies on an output gate, which is a neural layer of the output values computed by the LSTM units at the current time. C_t How many outputs are there to the LSTM at the present time. h_t

$$o_t = \sigma(w_1^0 \cdot x_t + w_h^0 \cdot h_{t-1} + b_0) \tag{7}$$

$$h_t = \sigma \cdot \tanh(C_t) \tag{8}$$

Where in f_t 、 i_t 、 o_t represents the state values of the forgetting gate, the memory gate, and the output gate, in that order; w_1^f 、 w_1^i 、 w_1^o 、 w_1^c is a weight matrix of the forgetting gate, the input gate, the output gate, and the tuple inputs connecting the inputs x_t to the components, in that order; w_A^f 、 w_A^i 、 w_A^o 、 w_A^c is a weight matrix of the forgetting gate, the input gate, and the output gate, and the tuple inputs bridging the previous layer of outputs h_{t-1} to the tuple; and b_f 、 b_i 、 b_o 、 b_c is a bias vector of the forgetting gate, the input gate, the output gate, and the tuple inputs.

ISSN 2616-5775 Vol. 6, Issue 10: 113-118, DOI: 10.25236/AJCIS.2023.061017

3. Results

3.1 Parameterization of the LSTM model

The first 70% of 1953 sets of data were divided into training set and the last 30% into test set to establish the LSTM air quality prediction model based on Kears framework. The AQI data are normalized and fed into the LSTM neural network for processing, and the LSTM-based AQI prediction model is obtained after several iterations.

When training the LSTM model [6], since the Adam algorithm can dynamically adjust the learning rate of each input parameter, the optimizer adopts the Adam optimization algorithm, and the absolute mean square error mae minimization is used as the optimization objective of the loss function. Since the output range of the Sigmoid function is between [0,1], and when the gradient values of the parameters are of the same sign, zigzag phenomenon is easy to occur when updating and it is difficult to find the optimal value. However, the tanh function takes values in the range of [-1,1], which avoids the shortcomings of the Sigmoid function, so we choose the activation function of the LSTM layer as the tanh function. In order to speed up the operation, a GPU was installed in the TensorFlow environment.

3.2 Analysis of model fitting results

Advance prediction of air quality can provide early warning information for relevant environmental protection departments, so that they can take appropriate measures to improve air quality, so this paper will take Beijing as an example to establish an air quality prediction model based on the long and short-term memory neural network (LSTM) using the daily average AQI data of Beijing for a total of 1,953 days from Jan. 1, 2018-April 30, 2023, for the purpose of establishing an air quality prediction model.

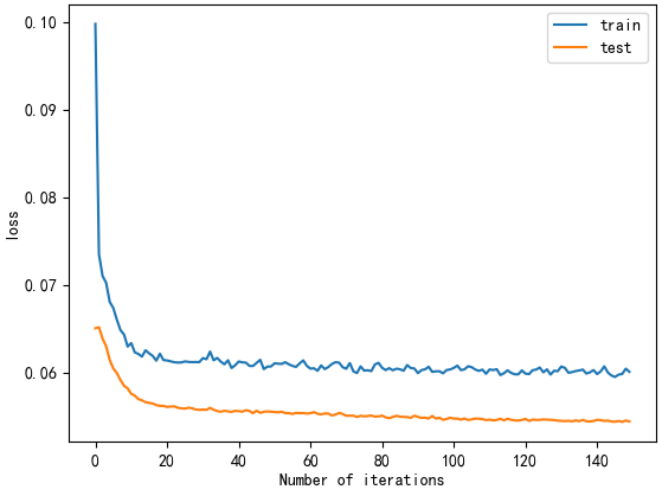

Figure 3: Model loss diagram

The model loss plot is shown in Figure 3, from the change of loss, we can see that the loss in the training set is rapidly decreasing and converging, which means that the model fits better on the training set. Similarly, in the test set, the loss of the model is also decreasing and finally stabilized, so it can be initially judged that the model fits better.

From the image, it can be seen that the trend of the fitted values and the true values fitted by the LSTM model are roughly the same, and the fitted values and the true values are basically on a line more often. Due to the fluctuation of the range of the values of the AQI in the same AQI class, such as 51-100 belonging to the same AQI class, the evaluation indexes have a large fluctuation. the same air quality class, so the large value of AQI can be understood as a normal situation , as shown in Figure 4.

ISSN 2616-5775 Vol. 6, Issue 10: 113-118, DOI: 10.25236/AJCIS.2023.061017

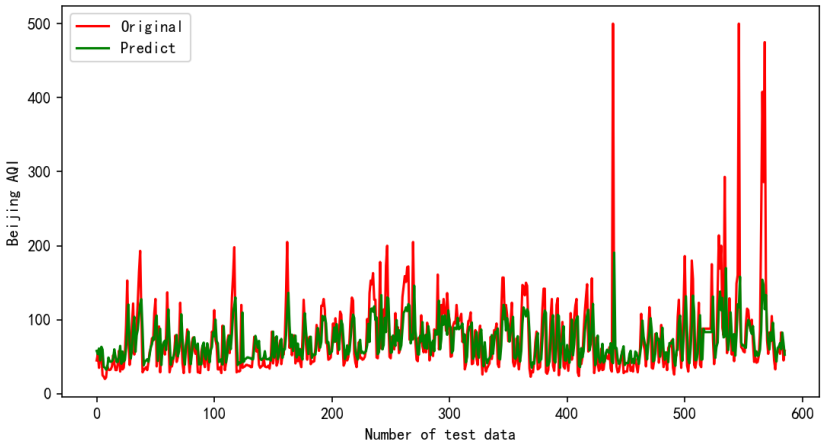

Figure 4: Plot of predicted versus true values of LSTM in the prediction set

In evaluating the prediction effect of the LSTM model, we simultaneously use the ARIMA model as the baseline model for comparison, and the evaluation indexes used are RMSE and MAE, and the specific experimental results are shown in the following Table 1.

Table 1: Comparison table of the prediction effect of the two models

Model Indicators	RESE	MAE
LSTM	44.847223	26.123211
ARIMA	26.483561	18.806126

Therefore, after comparing with the prediction accuracy of ARIMA model, it can be seen that although there is a gap between LSTM model and ARIMA model, the gap is not big, and it can be considered that the LSTM model performs well in the prediction of AQI, and it can be used to predict the air quality of the six cities.

3.3 Comparison of AQI Forecasts for Six Cities

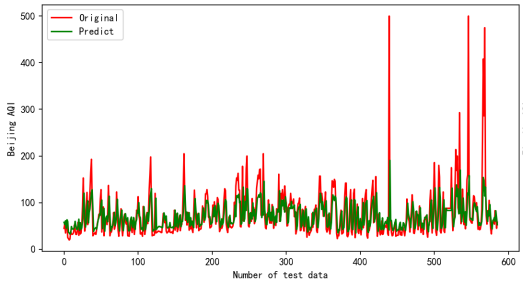

Figure 5: Beijing LSTM prediction map

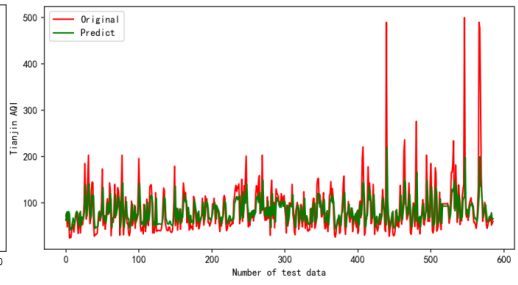

Figure 6: Tianjin LSTM prediction map

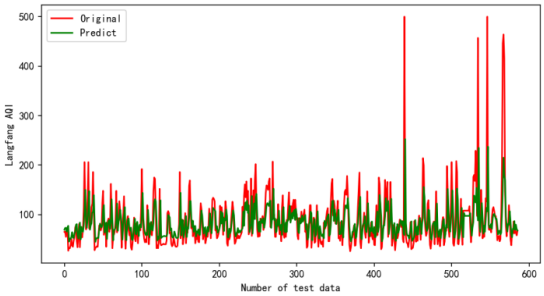

Figure 7: Langfang LSTM prediction map

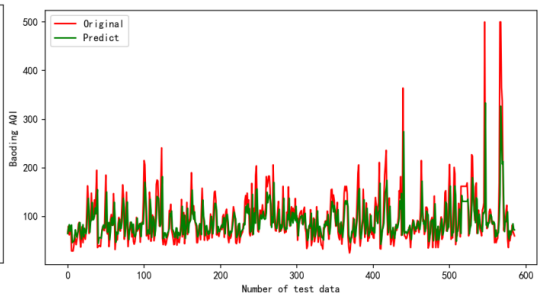

Figure 8: Baoding LSTM prediction map

ISSN 2616-5775 Vol. 6, Issue 10: 113-118, DOI: 10.25236/AJCIS.2023.061017

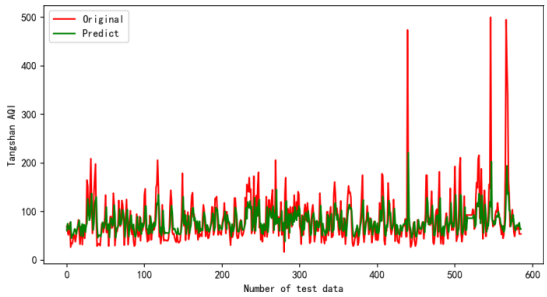

Figure 9: Tangshan LSTM prediction map

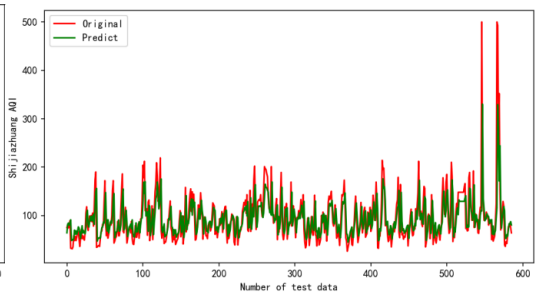

Figure 10: Shijiazhuang LSTM prediction map

Based on the long and short-term memory neural network model (LSTM), the AQI data of six cities, Beijing, Tianjin, Langfang, Baoding, Tangshan and Shijiazhuang, are predicted on the test set, and the prediction results are shown in Figures 5-10.

As can be seen from the figure, the volatility of the curves of the predicted and real values of AQI in each city is roughly the same, with a high degree of fit, reflecting that its error is within a certain range. The comparison of the regression evaluation indexes of the LSTM model is shown in the table 2 below, which is fluctuating within a certain reasonable range, so it can be assumed that the LSTM prediction model established in this paper has a good prediction effect in the dataset selected in this paper.

Table 2: Table of evaluation metrics for LSTM prediction models for six cities

	MSE	RMSE	MAE	R_square
Beijing	2011.273438	44.847223	26.123211	0.256921
Tianjin	2096.846924	45.791341	27.301731	0.241233
Langfang	2351.685547	48.494181	28.735516	0.233609
Baoding	1936.911011	44.010351	26.979553	0.329552
Tangshan	1955.81506	44.224598	26.902994	0.244060
Shijiazhuang	1644.649292	40.554276	25.264069	0.370502

4. Conclusions

The AQI is affected by six singularity factors, and the general linear regression model prediction is not accurate enough to analyze and predict the AQI, so this paper adopts the LSTM neural network prediction model for the AQI prediction. This model has a good self-assessment mechanism, and the LSTM model performs well in long-time prediction, which is more suitable for the dataset selected in this paper. From the evaluation indexes of this paper, we can see that the prediction effect of LSTM model is good, and it can improve the prediction accuracy of air quality, which is very helpful for air pollution monitoring, early warning and prevention. The Beijing-Tianjin-Hebei region, as one of the key development areas in China, is representative of the studies related to the air quality index in this region, which is of reference significance for the environmental management of other cities and is conducive to the realization of the goal of harmonious coexistence between human beings and nature. In the subsequent research, we will also consider more influencing factors comprehensively to improve the stability and accuracy of the model.

References

[1] Ai Meirong. Status and Prospect of Air Quality Evaluation Research [J]. Modern Business Industry, 2018(9): 188-189.
[2] Peng Sijun, Shen Jiachao, Zhu Xue. PM2.5 prediction based on ARIMA model [J]. Safety and Environmental Engineering, 2014, 32(6): 127-128.
[3] Wu L, Li N, Yang Y. Prediction of air quality indicators for the Beijing-Tianjin-Hebei region[J]. Journal of Cleaner Production, 2018, 196(pt.1-862):682-687.DOI:10.1016/j.jclepro.2018.06.068.
[4] He Zhiying, Hong Zhimin. Analysis of air quality conditions and influencing factors in Hohhot [J]. Journal of Inner Mongolia University of Technology (Natural Science Edition), 2021, 40(3): 190-198.
[5] Cao Tong, Bai Yanping. Research on air quality prediction by LSTM based on gradient descent optimization [J]. Journal of Shaanxi University of Science and Technology, 2020, 38(6): 159-164.
[6] Tang L , Zhou C , He L ,et al.Research on Air Quality of Beijing-Tianjin-Hebei Region based on SVM and Regression Analysis[C]//International Conference on Education.2017.DOI:10.2991/iceemr-17.2017.82.