Analysis and Prediction of Air Quality based on LSTM Neural Network —— Take Beijing Temple of Heaven as an Example

Hong Fang¹ and Lan Zhang¹
¹ 2360 Jinhai Road, Shanghai Pudong New Area, Shanghai Polytechnic University
Email: 153216496@qq.com

Abstract. LSTM is an excellent variant model of RNN, which inherits the characteristics of most RNN models and solves the problem of gradient disappearance caused by a gradual reduction in the process of gradient backpropagation. LSTM is very suitable for processing data highly related to time series. This paper selects meteorological elements and air quality data of the temple of heaven in Beijing from 2016 to 2018, analyse the correlation between meteorological elements and air quality, selects index modeling, and designs an air quality prediction model based on LSTM neural network. It is concluded that the model has better accuracy and robustness. Finally, reasonable air pollution control measures are proposed to provide a scientific theoretical basis and new prediction methods for the prevention and control of air pollution.

1. Introduction
In recent years, with the acceleration of industrialization and modernization in cities, the problem of air pollution has become increasingly prominent, especially in areas with a large number of people, which have seriously affected the daily production and life of residents. Therefore, the forecast for air quality index has been realized, and Taking preventive measures has important practical significance and social value. The Temple of Heaven is a world cultural heritage and a demonstration site of the national civilized scenic tourist area. Air pollution should be given more attention. To fundamentally solve the problem of air pollution, we must process and analyze real and accurate data, find the correlation between pollutants and air quality, establish an air quality prediction model, and take corresponding prevention measures, which is of great significance to optimize air quality and environmental protection.

Many scholars at home and abroad have made research and analysis of air quality and achieved many important results. Li Xiaofei et al. analyzed [1] the change characteristics of air quality in 42 Chinese cities from 2001 to 2010, and believed that there was a linear relationship between AQI and precipitation and wind speed. Wang Liyuan et al. took the air quality of Xichang city as an example[2], established a variable coefficient model, and used the local linear estimation method to fit the model to quantitatively analyze the change of the influence degree of meteorological factors in Xichang city on local air quality with the seasonal change. Forecasting is based on statistical methods and is based on statistics. It analyzes historical air data to predict future trends. With the continuous development of machine learning, it is widely used in the prediction of linear and non-linear and time series models. Common methods are mainly clustering methods[3], support vector machines[4-5], and neural networks[6]. Among many deep learning models, the recurrent neural network (RNN) introduces the concept of time-series into the design of the network structure, making it more adaptive in the analysis of time-series data. Among the many RNN variants, long-term and short-term memory[7] model
solves the problem of gradient disappearance and gradient explosion of RNN, so that the damaged neural network can effectively use long-term time series information.

2. Data Processing and Algorithm Overview

2.1. Data Sources
(1) The air quality index of the temple of heaven in Beijing was obtained from the air quality daily, published on the website of the ministry of environmental protection.

(2) Meteorological data are obtained from daily and monthly data of China.

2.2. Data Preprocessing
(1) Data set segmentation. Firstly, the air quality data should be divided into the training set and test set in a certain proportion. The training set is used for training model and training model parameters, and the test is used to detect the usability and accuracy of the model.

(2) Normalization of data[8]. In the field of machine learning, different evaluation indicators usually have different dimensions and dimension units, which will affect the analysis of data results.Normalization is to eliminate different dimensions of indicators. Data normalization is to scale the data so that it falls into a small specific interval, which is generally uniformly mapped to the interval [0,1], to eliminate the adverse effect caused by the singular sample data. Moreover, data normalization can improve the convergence speed and accuracy of the model. The representation of the unnormalized and normalized graphs is shown in figure 1.

![Figure 1. Unnormalized (left) and normalized (right).](image)

2.3. Kendall Rank Correlation Coefficient
Correlation analysis is a statistical method to study whether there is some kind of dependency relationship between phenomena and to explore the related direction and degree of correlation for a specific dependent phenomenon. When studying the degree of linear correlation between two variables, use the correlation coefficient to express. In statistics, the Kendall correlation coefficient[9] is named after Maurice Kendall, and its value is often expressed by the Greek letter \( \tau \) (tau). The Kendall correlation coefficient is a statistic used to measure the correlation between two random variables. A Kendall test is a no-parameter hypothesis test that uses calculated correlation coefficients to test the statistical dependence of two random variables. The Kendall correlation coefficient ranges from -1 to 1. When \( \tau \) is 1, it means that the two random variables have the same rank correlation; when \( \tau \) is -1, it means that the two random variables have completely opposite values. Rank correlation; when \( \tau \) is 0, it means that two random variables are independent of each other. Kendall correlation coefficient is a measure of the degree of relation between two ordered variables or two rank variables. Moreover, the influence of nodes (the rank order is the same) is taken into account. The correlation coefficient values range from -1 to 1. Kendall’s tau-b correlation coefficient is calculated as follows:
In this formula, $\tau$ is the Kendall's tau-b correlation coefficient of the variables $x$ and $y$; $x_i$ is the $j$-th observation of the variable $x$; $y_i$ is the $j$-th observation of the variable $y$; $t_i$ is the number of nodes in the $i$-th group of the variable $x$; $\mu_i$ is the number of nodes in the $y$-th group of the variable $y$; $\phi$ is the function independent variable.

\begin{equation}
\tau = \frac{\sum_{j=1}^{i} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)}{\sqrt{\frac{n(n-2)}{2} \sum_{j=1}^{i} t_j (t_j - 1)}}
\end{equation}

\begin{equation}
\text{sgn}(\phi) = \begin{cases} 
1, & \phi > 0 \\
0, & \phi = 0 \\
-1, & \phi < 0 
\end{cases}
\end{equation}

2.4. LSTM Algorithm Principle

Recurrent neural network (RNN) is a powerful deep neural network that has significant effects on long-term dependent time-series data processing. To solve the problem of vanishing gradients when dealing with long-term dependence, Hochreiter & Schmidhuber proposed a long-term and short-term memory network model. Compared with traditional RNNs, LSTM (long-term and short-term memory networks) has a more sophisticated information transfer mechanism, which can effectively solve long-term data Dependence issues. At the same time, as the basic subunit of the encoder-decoder framework, LSTM can also realize the encoding and decoding of time series data, and replace the hidden layer neurons in RNN with memory units to realize the memory of past information, and each memory unit contains one or more memory cells and three-door controllers. The LSTM structure is shown in figure 2.

![Figure 2. LSTM structure diagram.](image)

![Figure 3. Schematic diagram of LSTM neural network structure.](image)

The basic structure of LSTM and RNN is very similar, with a chain structure. But LSTM has a very useful mechanism, namely forgetting gate. In LSTM, not only a single network-layer is used, but four modules interact with each other in a special way, as shown in figure 3.

3. The Experimental Process

Kendall correlation coefficient was used to analyze the relationship between AQI and weather conditions, maximum and low temperature, minimum air temperature, average air temperature, and average wind speed, and AQI was numerically processed. First, the weather conditions were
reclassified and expressed as seven weather features, and the processed data were expressed as 1-7, as shown in table 1.

Table 1. Data preprocessing.

| Raw data                        | Data conversion |
|---------------------------------|-----------------|
| sunny day                       | 1               |
| partly cloudy, smog             | 2               |
| cloudy day                      | 3               |
| Light rain, shower, thunder     | 4               |
| Rain, heavy rain, rainstorm     | 5               |
| Light snow, Medium snow         | 6               |
| Floating dust, Dust             | 7               |

Table 2. Correlation analysis of meteorological elements and AQI.

| AQI                         | The weather conditions | The highest temperature | The lowest temperature | The average temperature | Average wind speed |
|-----------------------------|------------------------|-------------------------|------------------------|-------------------------|--------------------|
| Correlation coefficient     | 1.000                  | .061**                  | .101**                 | .091**                  | .096**             | -0.055*            |
| Significance (double-tailed)|                        |                         |                        |                         |                    |                    |
| N                           | 1095                   | 1095                    | 1095                   | 1095                    | 1095               | .014               |
| Correlation coefficient     | .061**                 | 1.000                   | .096**                 | .166**                  | .130**             | .026               |
| Significance (double-tailed)|                        |                         |                        |                         |                    |                    |
| N                           | 1095                   | 1095                    | 1095                   | 1095                    | 1095               | .300               |
| Kendall tau_b               |                        |                         |                        |                         |                    |                    |
| Correlation coefficient     | .101**                 | .096**                  | 1.000                  | .880**                  | .947**             | .023               |
| Significance (double-tailed)|                        |                         |                        |                         |                    |                    |
| N                           | 1095                   | 1095                    | 1095                   | 1095                    | 1095               | .294               |
| Correlation coefficient     | .091**                 | .166**                  | .880**                 | 1.000                   | .945**             | .014               |
| Significance (double-tailed)|                        |                         |                        |                         |                    |                    |
| N                           | 1095                   | 1095                    | 1095                   | 1095                    | 1095               | .533               |
| Correlation coefficient     | .000                   | .000                    | .000                   | .000                    | .000               | .408               |
| Significance (double-tailed)|                        |                         |                        |                         |                    |                    |
| N                           | 1095                   | 1095                    | 1095                   | 1095                    | 1095               | .108               |
| Correlation coefficient     | .096**                 | .130**                  | .947**                 | .945**                  | 1.000              | .018               |
| Significance (double-tailed)|                        |                         |                        |                         |                    |                    |
| N                           | 1095                   | 1095                    | 1095                   | 1095                    | 1095               | .300               |
| Correlation coefficient     | .000                   | .000                    | .000                   | .000                    | .000               | .408               |
| Significance (double-tailed)|                        |                         |                        |                         |                    |                    |
| N                           | 1095                   | 1095                    | 1095                   | 1095                    | 1095               | .108               |
| Correlation coefficient     | .096**                 | .130**                  | .947**                 | .945**                  | 1.000              | .018               |
| Significance (double-tailed)|                        |                         |                        |                         |                    |                    |
| N                           | 1095                   | 1095                    | 1095                   | 1095                    | 1095               | .300               |

**. The correlation was significant above 0.01 (double-tailed)
*. The correlation was significant above 0.05 (double-tailed)

As shown in table 2, the Kendall coefficient was used to analyze the correlation between AQI and meteorological elements and the double-tailed significance. According to the comparison and analysis of correlation coefficients, three variables including weather conditions, average temperature, and average wind speed were selected as input values of the prediction model in the subsequent modeling.
The development environment of this experiment is PyCharm 2019.2.3, the machine learning framework is Keras, Keras is a deep learning library based on TensorFlow and Theano[10], and Keras is a high-level neural network API completely edited by Python and can only be used for the development of Python.

Because there are missing values in the crawled air data, to avoid the impact of missing values on the prediction of the data, compared with other interpolation methods, random forest interpolation is used in this article [11]. This method can effectively process high-dimensional data, effectively extract auxiliary variable information, and is suitable for processing missing data in the context of big data. In this paper, 40% of the data is used to train the model, and 40% is used as a test set to verify the model. The model can predict air quality. The LSTM prediction data is shown in Figure 4. A common prediction model is the BP neural network, but the BP is essentially a gradient descent method, with slow learning speed and poor stability. The BP prediction model is shown in Figure 5.

Because the image can only roughly see the fitting trend between the predicted value and the true value, it is greatly affected by subjective factors. Therefore, we need to use a scientific metric to evaluate the performance of the prediction. We used root mean square error (standard formula), root mean square error, and prediction accuracy to evaluate the predicted results, as shown in table 3. The LSTM prediction has a lower error rate and shorter time, making it more suitable for air quality data prediction.

| Test RMSE(standardized) | Test RMSE | Prerate       |
|-------------------------|-----------|---------------|
| LSTM                    | 0.031     | 15.770        |
| BP                      | 0.033     | 16.994        |

4. Conclusion
To sum up, in this experiment, LSTM is more suitable for air quality prediction than BP. The fifth plenary session of the 18th CPC central committee included ecological environment protection in the agenda of green development. Big data technology can integrate, process and analyze massive air data, and this technology that breaks through traditional information processing can bring new opportunities to the joint prevention, control and supervision of air pollution [12].

1. Analyze air quality monitoring data based on big data
At present, air pollution monitoring mainly focuses on the monitoring and control of a single pollutant, including nitrogen oxides, greenhouse gases and sulfides. However, it is unable to carry out coordinated monitoring and analysis of multiple pollutants. However, in real life, a variety of
pollutants often appear at the same time and related reactions take place, leading to more serious air pollution events. The emergence of big data technology effectively solves the existing problems of current air monitoring. Big data technology can cooperate to analyze monitoring videos, pictures, texts and other non-structural data to understand the regional scope of air pollution, sources of pollution and health risks brought by pollutants, providing support for the management department to prevent and control air pollution.

2. Information sharing with other industries through big data
   At present, mainly by the environmental protection departments to carry out the air pollution in the information management work, has not fulfilled with meteorological data sharing, transportation and other departments of industry, in order to improve the quality of air environment management, environment management department can through the comprehensive analysis of forestry, the technology of data traffic, water conservancy and other industries more than information data, from different angles, the early warning system for environmental air quality forecast that can enhance the effect of air pollution monitoring, scientific planning, air pollution prevention and control measures[13].

3. Big data provides support for decision analysis
   At the present stage, China's air quality management is still a typical territoriality model, with different territoriality and departments at all levels separated from each other and lack of communication, making it impossible to coordinate the treatment of air pollution [14]. With the progress of the times and the innovation of ideas, the Beijing-Tianjin-Hebei region, the Yangtze river delta region and other developed city clusters have begun to carry out the joint prevention and control of regional air pollution, but these good experiences and practices have not been well promoted. Regional air pollution zone spreading can use big data technology, comprehensive analysis of the regional government management level, economic development and status quo elements such as air pollution, in the clear air quality objectives and the maximum keep balance under the condition of the various elements of scientific decisions, in order to establish the cooperation mechanism that can meet the needs of each region, building common air pollution zone spreading regulatory system.

5. Acknowledgments
   This paper is funded by the program of the key discipline “Applied Mathematics” of Shanghai Polytechnic University(XXKPY1604).

6. References
   [1] Li Xiaofei, Zhang Mingjun, Wang Shengjie, Zhao Aifang, Ma Qian. Analysis of the characteristics and influencing factors of China's air pollution index[J]. Environmental Science, 2012, 33 (06): 1936-1943.
   [2] Wang Liyuan, Li Yuqi. Analysis of Meteorological Factors Affecting Air Quality in Xichang City——A Study Based on Variable Coefficient Model[J]. Journal of Yangtze Normal University, 2014, 33 (05): 98-102 + 112.
   [3] Cevik H H, Cunkas M. Short-term load forecasting using fuzzy logic and ANFIS. Neural Comput Appl, 2015, 26 (1), 1355–1367.
   [4] Li P, Tan Z, Lili Y, et al. Time Series prediction of mining subsidence based on a SVM. International Journal of Mining Science and Technology, 2011, 21 (4): 557-562.
   [5] Pai P F, Hong W C. Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms. Electric Power Systems Research, 2005, 74 (3): 417-425.
   [6] Yu Weizhong, Chen Jian. Application of BP artificial neural network model in forecasting urban air pollution [J]. Pollution Control Technology, 2013, 26 (3): 55-57.
   [7] GRAVES A. Long short-term memory[M]. Berlin: Springer, 2012: 1735-1780.
   [8] Gao Min, Qiu Tianlei, Jia Ruizhi, Han Meilin, Song Yuan, Wang Xuming. Characteristics of Bioaerosol Concentration and Particle Size in Beijing Haze Weather[J]. Environmental Science, 2014, 35 (12): 4415-4421.
   [9] Wang Qinde, Zhao Yi. Several calculation methods and principles of the Kendall coefficient [J].Journal of Shanxi Agricultural University, 1993 (03): 225-228.
[10] Zeng Hao. Research on the prediction model of environmental pollution time series based on LSTM[D]. Huazhong University of Science and Technology, 2019.

[11] Meng Jie, Li Chunlin. Interpolation of Missing Values of Classified Data Based on Random Forest Model[J]. Forum on Statistics and Information, 2014, 29 (09): 86-90.

[12] Zheng Shiming, Liu Jiajun. Air Pollution Control and Government Decision Based on Big Data [J]. Journal of South China Normal University (Social Science Edition), 2017 (04): 105-111 + 1.

[13] Han Jin. Research on some key technologies of big data service. Doctoral dissertation of Beijing university of posts and telecommunications, 2013.

[14] Wang Shuwei. Research on information resource sharing strategies among government departments in the era of big data. Master's thesis of Jilin University, 2013.