Research on the air quality prediction model of Wuhai mining area based on deep learning

Jinghua Wang[1], *Jin* Cheng[2], *Fang* Liu[1], *Lei* Yan[1,*], *and Taijie* Tang[1]

[1]School of Technology, Beijing Forestry University, Beijing100083, China
[2]College of Biological Sciences and Technology, Beijing Forestry University, Beijing100083, China

Abstract. With the large-scale and high-intensity mining of coal resources in the Wuhai mining area, the destruction of soil and erosion of rocks has intensified, causing a large amount of surface soil spalling from the mine body and serious damage to the surface vegetation, which has had a serious impact on the quality of the environment in and around the mine. This paper focuses on the corresponding early warning research on air quality in the mining area of Wuhai, and constructs Deep Recurrent Neural Network (DRNN) and Deep Long Short Time Memory Neural Network (DLSTM) air quality prediction models based on the filtered weather factors. The simulation results are also compared and find that the prediction results of DLSTM are better than those of DRNN, with a prediction accuracy of 92.85%. The model is able to accurately predict the values and trends of various air pollutant concentrations in the mining area of Wuhai.

1 Introduction

Wuhai is one of the important coal production bases in China, located in the southwest of Inner Mongolia and the western border of the Ordos massif. The city contains a large amount of coking coal underground, with reserves occupying 58.8% of the total amount in the region.However, in recent years, with the vigorous exploitation of mineral resources, the ecological environment of Wuhai has been severely damaged and the situation has deteriorated rapidly.

2 Data sources

In this paper, the weather data of Wuhai area are obtained by the method of reptile, including air quality data and meteorological data.Data from www.pm25.in / wuhai.Using hourly monitoring data for the period from 1 January 2015 to 4 December 2019, for each point in time, three stations collect data simultaneously, that is, 24 times a day, with a total of 72 data collected by the three sites.

* Corresponding author: mark_yanlei@bjfu.edu.cn

3 Data preprocessing and prediction factor selection

3.1 Data standardization

The initial data will be standardized to avoid a computational failure due to abnormal convergence of the model caused by an unstandardized dataset. According to the unclear distribution of the original data, Z-SCORE is chosen to standardize the raw data.

The principle of Z-SCORE standardization is to calculate and derive the average value and mean value of the data, and the value of the interval near the zero value can be used to represent the original data value. Analyze the processed data through (0, 1) standard normal distribution. Such as formula (1):

$$Y(x)=\frac{x-\bar{x}}{\alpha} \tag{1}$$

$\bar{x}$: average difference of all elements α : Standard deviation

3.2 Correlation analysis and prediction factor selection

In this paper, the principal component analysis is used to select prediction factors from a large number of data items and select those that have a greater impact on air pollution concentration.We also pay attention to seasonal effects on air quality, and ensure the accuracy of air-quality prediction models. Select the data values from 00:00 on July 4, 2018 to 00:00 on July 10, including air quality data such as $PM_{2.5}$, O3, CO, NO_2, PM_{10}, and SO_2 concentrations.

Use MATLAB R2014a to analyze air quality data and meteorological data. Correlation values are shown in table 1:

Table 1. PM2.5 Correlation coefficient table.

Correlation coefficient	Highest temperature	Lowest temperature	Wind power	O_3	CO	NO_2	PM_{10}	SO_2
Winter	0.29	-0.01	-0.25	-0.35	-0.13	0.48	0.78	0.47
Summer	0.37	-0.05	-0.23	-0.38	-0.56	0.66	0.94	0.37

The correlation between the data items is determined by the correlation coefficient. As shown in Table 1, taking $PM_{2.5}$ as an example, the correlation between $PM_{2.5}$ and meteorological factors is low, and it is significant with data items such as CO, SO_2, NO_2, and O_3 Correlation, and PM_{10} belong to high correlation. By calculating the correlation coefficient, PM_{10}, $PM_{2.5}$, O_3, CO, NO_2, and SO_2 are finally selected as the input data of the prediction model.

4 Air quality prediction model based on DRNN

4.1 Deep recurrent neural network (DRNN)

Deep Recurrent Neural Network is one of the deep learning algorithms, and its main function is to construct a network of sequence data.

4.2 Simulation results and analysis

Select the data of March 1, 2019 as the test set, input the trained Deep Recurrent Neural Network prediction model, and then compare the predicted value of the model with the true value. The result is shown in Figure 1-6.

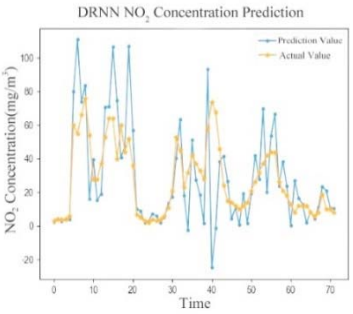

Fig. 1. Simulation comparison of NO_2.

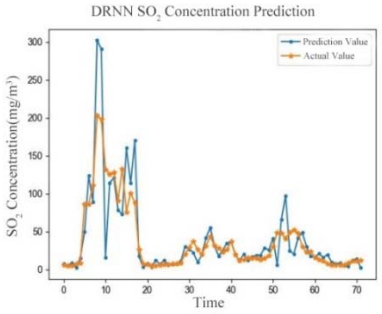

Fig. 2. Simulation comparison of SO_2.

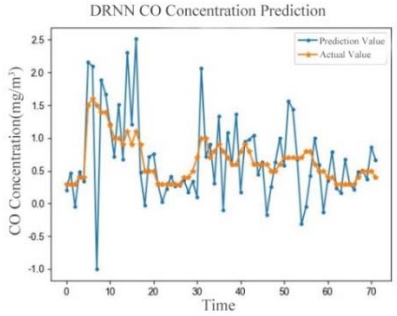

Fig. 3. Simulation comparison of CO.

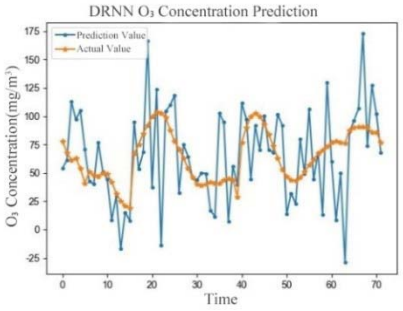

Fig. 4. Simulation comparison of O_3.

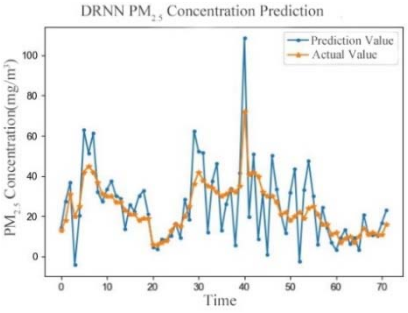

Fig. 5. Simulation comparison of $PM_{2.5}$.

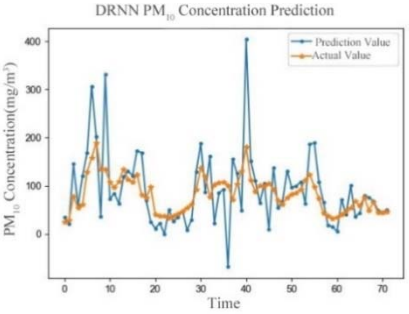

Fig. 6. Simulation comparison of PM_{10}.

The predicted experimental results of the deep recurrent neural network air quality forecasting model for each forecasting project are recorded as shown in Table 2.

Table 2. Experimental record of prediction model based on deep circulation neural network.

Number of hidden layers	Number of hidden layer units	Excellent	Accept.	Not accept
1	10	82.33%	20.13%	0.28%
2	10 9	83.15%	19.87%	0.28%
3	10 9 8	83.76%	19.56%	0.28%
4	10 9 8 7	84.21%	19.24%	0.00%
5	10 9 8 7 6	84.53%	18.87%	0.00%
6	10 9 8 7 6 5	85.42%	17.64%	0.00%
7	10 9 8 7 6 5 4	86.79%	16.36%	0.00%
8	10 9 8 7 6 5 4 3	88.70%	15.96%	0.00%

Excellent: How well the model predicts the results Accept: Acceptability of model prediction results Not accept: Unacceptable degree of model prediction results

It can be seen from Table 2 that the prediction accuracy of the model is affected by the number of hidden layers when the deep recurrent neural network is used to predict the nodes. Table 2 shows the record of the prediction accuracy of different hidden layers. When the number of hidden layers is 8, the accuracy of the prediction model is the highest, reaching 88.70%.

5 Air quality prediction model based on DLSTM

5.1 Deep long and short time memory neural network (DLSTM)

Although the prediction model based on the recurrent neural network performs better in air quality prediction, the processing efficiency of the prediction model based on the recurrent neural network begins to decline when encountering data with a longer output sequence. To avoid this problem, LSTM was chosen to build the air quality prediction model.

5.2 Simulation results and analysis

Select the data of March 1, 2019 as the test set, input the trained deep and short-term memory neural network prediction model, and then compare the predicted value of the model with the true value. The result is shown in Figure 7-12.

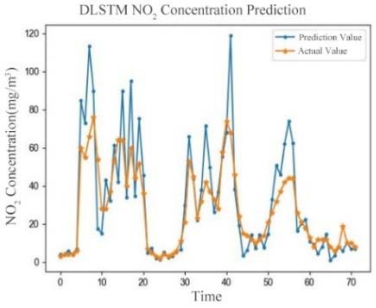

Fig. 7. Simulation comparison of NO_2.

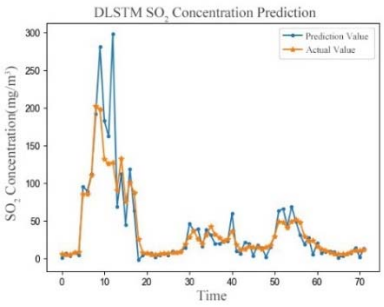

Fig. 8. Simulation comparison of SO_2.

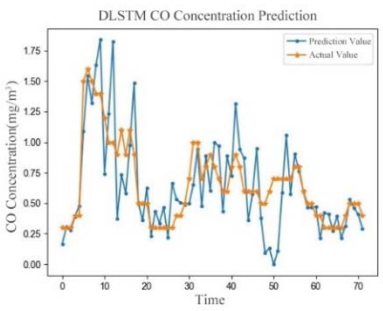

Fig. 9. Simulation comparison of CO.

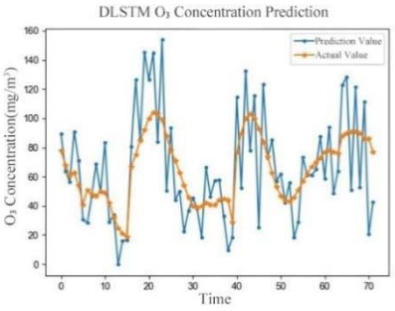

Fig. 10. Simulation comparison of NO_2.

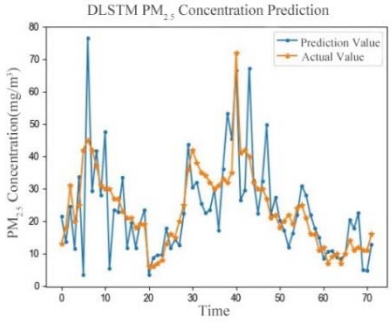

Fig. 11. Simulation comparison of $PM_{2.5}$.

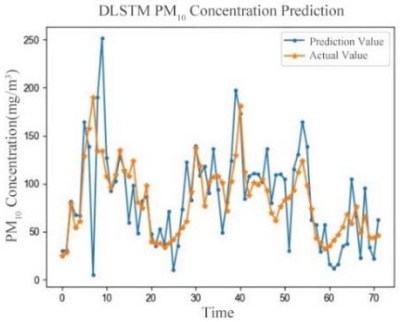

Fig. 12. Simulation comparison of PM_{10}.

The results of the prediction experiment results of each prediction item of the deep long and short time memory neural network air quality prediction model are shown in Table 3:

Table 3. Prediction results of neural network model based on DLSTM.

Number of hidden layers	Number of hidden layer units	Excellent	Accept.	Not accept
1	10	83.18%	25.23%	0.28%
2	10 9	84.52%	22.65%	0.28%
3	10 9 8	85.83%	21.55%	0.28%
4	10 9 8 7	87.92%	20.83%	0.00%
5	10 9 8 7 6	87.27%	19.27%	0.00%
6	10 9 8 7 6 5	88.38%	17.64%	0.00%
7	10 9 8 7 6 5 4	92.85%	15.96%	0.00%
8	10 9 8 7 6 5 4 3	89.49%	16.36%	0.00%

Excellent: How well the model predicts the results Accept: Acceptability of model prediction results Not accept: Unacceptable degree of model prediction results

The above table shows that the accuracy of LSTM based prediction model is affected by the number of hidden layer nodes. When the number of hidden layers is 7, the prediction results is the best, and the accuracy is 92.85%. Compared with the air quality prediction model based on deep recurrent neural network, the air quality prediction model based on deep LSTM can more accurately predict AQI, SO_2, CO, NO_2, O_3, PM_{10}, $PM_{2.5}$ concentration Equivalent, achieved the expected effect.

6 Conclusions

This paper uses DRNN and DLSTM to build an air quality prediction model, and conducts

experiments on the collected data. The experimental results show that the accuracy of the air quality prediction model based on DLSTM is 92.85% higher than the accuracy of the air quality prediction model based on DRNN, which is 88.70%. Forecast of air quality and its changing trend in Wuhai mining area. This research result can provide powerful technical support for ecological security in Wuhai mining area, and it is very important to standardize the production and accelerate the ecological recovery of mining areas.

Acknowledgments

Thanks for the funding of national key research and development project "Ecology Safety Assurance Technology of Coal Base in Northwest Arid Desert Area" (2017YFC0504403) and all members of the project team for their strong support during the research process. At the same time, I sincerely thank the experts and editors for their valuable comments on this paper during the review process.

References

1. L. Amirhajlou, Z. Sohrabi, M. Alebouyeh, N. Tavakoli, R. Haghighi,A. Hashemi, A. Asoodeh. Journal of Education and Health Promotion, **8**(1). (2019)
2. Bo Z., Jiashi F., Xiao W., Shuicheng Y. International Journal of Automation and Computing, **14**(02):119-135. (2015)
3. Wang,G. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, **40**(8). (2018)
4. Wei-Jie G., Xue-Yan Z., Kian Fan C., Nan-Shan Z. The Lancet, **388**(10054). (2016)
5. Ialongo C., Pieri M., Bernardini S. Clinical chemistry and laboratory medicine,**55**(2). (2017)
6. K.M.Mok, K.I.Hoi, Giuseppe. Future Generation Computer Systems, **25**(96):731-749 (2005)
7. Mohammad H.A., Alireza B., Mahyar G., Masoud H., Roghayeh G., Mohammad R. N. Journal of Thermal Analysis and Calorimetry,**139**(3). (2020)
8. Mohammad A., Ashish S., Kumari K., Jamshed A. Radio electronics and Communications Systems,**62**(8). (2016)
9. Maher E., Amir H. M., Mohammad S.T. International Journal of Greenhouse Gas Control,**58**. (2017)
10. Man W., Tsan Y., Esmond M. Sensors,**14**(11). (2014)
11. Nematollahi M., Akbari R., Nikeghbalian S., Salehnasab C. International journal of organ transplantation me dicine,**8**(2). (2017)
12. Soleiman H., Shahin R., Mortaza A., Seyed S. M. Drying Technology,**33**(2). (2015)
13. Alireza B., Alireza B., Amir H.M., Amirreza B. International Journal of Greenhouse Gas Control,**57**. (2016)
14. Alireza B., Mohammad A., Bahram H. S. The Journal of Supercritical Fluids,**98**.(2015)