Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
e-ISSN: 2395-3454

Vol. 19 | No. 1 | pp. 229-236 | 2020

Original Research Paper | Open Access

LSTM-based Air Quality Predicted Model for Large Cities in China

Shuyue Zhang*, Minfeng Lin*, Xiuguo Zou*†, Steven Su, Wentian Zhang**, Xuhui Zhang* and Zijie Guo***
*College of Engineering, Nanjing Agricultural University, Nanjing, Jiangsu, 210031, China
**Faculty of Engineering and Information Technology, University of Technology Sydney, NSW 2007, Australia
†Corresponding author: Xiuguo Zou

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 29-04-2019
Accepted: 02-07-2019

Key Words:
Air Quality
LSTM model
$PM_{2.5}$

ABSTRACT

In this paper, the LSTM model is used to predict the PM2.5 concentrations in five representative Chinese cities with the GDP exceeding 1 trillion Yuan, including Beijing, Chengdu, Shanghai, Shenzhen and Wuhan. The $PM_{2.5}$ concentration data in 2015-2017 are selected for training, and the results are optimized to achieve an efficient solution by adjusting the parameters. Based on the optimized solution, a test is carried out to predict the $PM_{2.5}$ concentration in 2018, and the results are compared with the real value obtained from the monitoring centre. According to the comparison results, the correlation coefficient of Wuhan and Chengdu is 0.86724 and 0.80070, which are the highest in these five cities. While the correlation coefficient of Shenzhen and Shanghai, are 0.78225, 0.72147, Beijing, as the capital city of China achieved the lowest correlation coefficient which is 0.64118. The LSTM-based predictive model has relatively good reliability and transferability. More effective predictive results can be achieved by implementing deep learning to analyse PM2.5 concentration.

INTRODUCTION

Due to the speeding up of the industrialization, environmental pollution issues have rapidly escalated since the 20th century. Particularly, the air pollution caused by $PM_{2.5}$ becomes more and more prominent. According to the findings by research scholars, $PM_{2.5}$, compared with other pollutants, contains more substances that damage human health, especially for the respiratory system. Moreover, $PM_{2.5}$ also has a severe effect on the visibility of atmospheric, and climate change. Therefore, $PM_{2.5}$ becomes a key research topic in the field of environmental science. Based on the assessment of U.S. Environmental Protection Agency (EPA), the economic loss caused by death accounted for 89% of the total loss caused by the health losses related to $PM_{2.5}$ pollution (Evans et al. 2019). In China, considering the research methods of $PM_{2.5}$ concentration is restricted by many factors (e.g. the vast territories, complex topographic changes, variable meteorological conditions, and large populations, etc.), general research methods like projection pursuit regression are not suited to solve $PM_{2.5}$ issues (Liu et al. 2019). Therefore, the methods of multivariate statistical analysis and traditional machine learning have been chosen by most scholars. In 2016, Fu et al. (2016) proposed five factors contributing to the smog pollution in Jiangsu Province based on the data collected from 2000 to 2014. By using the multiple linear regression model, a quantitative measure of the influence degree and significance level of these factors are studied in this paper. This paper also researched the specific measures for smog management from factors with significant influence. In the same year, He et al. (2016) analysed the spatial distribution features of $PM_{2.5}$ concentration in 2014 in Jiangsu Province using Kriging interpolation method. A grey correlation model is used to calculate the degree of association between $PM_{2.5}$ concentration and influence factors and shown its superiority compared with the other methods. Therefore, the method of multivariate statistical analysis plays an important role in $PM_{2.5}$ concentration analysis. However, it is not possible to use this method to obtain a good fitting model for air pollution prediction and to cope with sudden environmental changes. As a result, it has been mostly used in the auxiliary study of atmospheric pollution in recent years.

George et al. (2000) used the traditional machine learning method Bayesian Maximum Entropy (BME) mapping to study the influence factors of $PM_{2.5}$ concentration. Wang et al. (2009) explored a new urban air pollution prediction model based on Bayesian network, which is mainly designed for the middle and small cities in China. Zhou et al. (2004) proposed a BP network model for predicting air pollution indexes and researched comparative prediction. According to the verification result, the predictive model generated based on the BP method can provide an accurate prediction of mutational trends and is better than the stepwise regression model. Bai et al. (2013) researched the ten years (2001 to 2010) surface meteorological observation data and air pollution indexes in Beijing using the BP neural network model. In this paper, the air pollution index predictive model which trained in different seasons shows the superior performance of the BP model.

Relatively high accuracy can be obtained in using the machine learning method to predict atmospheric pollutant concentration, but how to combine the various attributes of nodes more effectively and achieve the better performance is that we need to do. The traditional machine learning methods have very poor transferability and limited application. Since 2012, the deep learning method has attracted considerable attention all over the world and stepped into a booming development stage. Deep learning has given rise to many frameworks which can be applied to various practical problems. Besides, the performance of the deep learning model is rapidly improved and its accuracy in dig data prediction is far beyond other models. The algorithms for deep learning are regenerated constantly, therefore the problems in $PM_{2.5}$ concentration analysis can be well solved with the development of the more effective models. Yin et al. (2015) used the deep learning framework DBN to establish an air pollution predictive model, through which they found out that the predictability of this model is higher than other classical predictive models under the two evaluation indexes set by them. Sun (2018) predicted the air quality indexes using the deep belief network predictive model, where the accuracy of predicted results is 91.4%, which is 32.2% and 25.7% higher than that of the integrated auto-regressive moving average model and BP neural network model. The LSTM model has been widely used by scholars considering its long-term and short-term memory function and forgetting mechanism. Li et al. (2018) designed an experiment based on the power plant load data and found out that the predicted effect of LSTM recurrent neural network algorithm based on Tensor Flow is better than that of the traditional machine learning algorithms. The model showed good robustness with the increasing number of data. An et al. (2018) established an LSTM model and analysed the degree of fitting and accuracy of the model, providing a reference for tilapia breeding and cultivation. Besides, LSTM also has good effects on practical problems such as short-term traffic flow (Qiao et al. 2017), stock price trend (Liu 2016), ship track (Quan et al. 2018) and other data trends that are difficult to determine.

In this paper, the original features of $PM_{2.5}$ concentration data can be well turned into a new feature space by using the LSTM algorithm. This method can obtain a more hierarchical feature representation through active learning, which can improve the predicted performance. Compared with the traditional methods, this method is paid more attention to the change rules of $PM_{2.5}$ concentration. In addition, according to the results obtained with other existing air pollutant predicted methods, the appropriate application range and advantages and disadvantages of models could be learned through active learning.

MATERIALS AND METHODS

Study Objects

There are 14 cities with GDP reaching 1000 billion Yuan in China, and all these cities are in the southeast of “Hu Line”, which makes “Hu Line” become the dividing line of urbanization level to a certain extent. “Hu Line” is a boundary dividing the population density in China by “Aihui-Tengchong” Line (from Heihe City of Heilongjiang Province to Tengchong City of Yunnan Province) proposed by Chinese geographer Huanyong Hu in 1935 (Guo et al. 2016). Pursuant to the fifth national population census in 2000, the calculation by using ArcGIS software shows that the south-eastern region accounts for 43.8% of the total national area and the population accounts for 94.1% according to the “Hu Line”. It suggests great significance in economic production, social development, and scientific research.

The spatial distribution of the 14 cities is shown in Fig. 1, where the city locations are represented by blue dots, and the straight line represents “Hu Line”. The $PM_{2.5}$ concentration data at 3-9 monitoring points of each city has been recorded.

Five representative cities, including Shanghai, Shenzhen, Chengdu, Beijing, and Wuhan, are selected from these 14 cities as the study objects for air quality analysis. The selected cities are separate in the East of China, the South of China, the West of China, the North of China, and Central China respectively. All of them are in the southeast of “Hu Line, and located in East China, South China, West China, North China, and Central China respectively.

Data Acquisition

China National Environmental Monitoring Centre releases the monitoring concentration data of $PM_{2.5}$ and other standard pollutants every hour in Shanghai, Shenzhen, Chengdu, Beijing and Wuhan. The $PM_{2.5}$ concentration data at 3-9 monitoring points of each city has been recorded every one hour. We have downloaded the data of each point from January 2015 to December 2018. The statistical analysis results of $PM_{2.5}$ data are shown in Table 1.

There are some unavoidable errors in the large sample test. Improving data quality is an effective way to reduce errors. For this reason, null values in the measurement are eliminated. The table in the above shows the result after abnormal data is eliminated. From the statistical analysis of $PM_{2.5}$ concentration data combined with the urban spatial distribution map, it can be found that while Beijing has

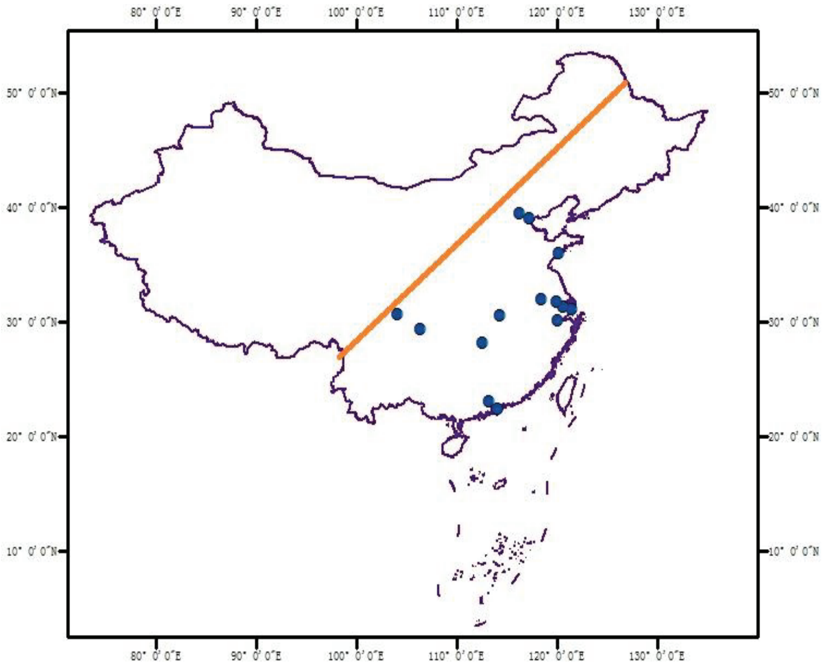

Fig. 1: The spatial distribution of the 14 cities with "Hu Line".

Table 1: The statistical analysis results of $PM_{2.5}$ data.

City	Data Volume	Mean value	Standard Deviation	Maximum value	Minimum value
Beijing	339119	63.27370	56.62778	457.97570	4.32110
Shanghai	287847	41.47401	25.61609	204.22689	5.64216
Wuhan	297741	57.22271	35.14301	280.37768	8.24057
Shenzhen	358087	27.70851	14.34911	100.72137	6.56870
Chengdu	260548	55.46151	37.24526	290.30208	7.26984

the largest $PM_{2.5}$ mean value, the smallest $PM_{2.5}$ mean value is recorded in Shenzhen. The $PM_{2.5}$ concentration in north-western China is higher than that in the southeast.

Data Processing Method

Data normalization: Data normalization is a processing method of transforming dimensional data into non-dimensional data. In order to improve the rate of convergence and the accuracy of the models, the data were normalized.

The purpose of data normalization is to concentrate all values into the interval of (0,1). The commonly-used data normalization methods are linear function normalization, and 0 mean standardization. The 0 mean standardization method is usually used when the data distribution approximates to Gaussian distribution. It is obvious that the data in the study does not meet this condition. Therefore, we adopted linear function normalization. The normalization equation is shown in Equation (1).

$$X_{norm} = \frac{X}{X_{\max} - X_{\min}} \quad \ldots(1)$$

Where, X_{norm} represents normalized data, X represents original data, X_{max} represents the maximum in original data, and X_{min} represents the minimum in original data.

LSTM model: The LSTM model, namely long short-term memory model, was adopted in the study. In LSTM, a node like the valve is added to the RNN structure. There are three kinds of valves: forgetting valve, input valve, and output valve. According to the on-off status of these valves, it can be determined that whether the network's memory state has reached the threshold in the output results of the layer and then is added to the calculation of the layer (Graves et al. 2012). RNN is still a neural network. The recurrent neural network can be sued to find the sequence correlation between samples. As a result, it is often used for modelling of

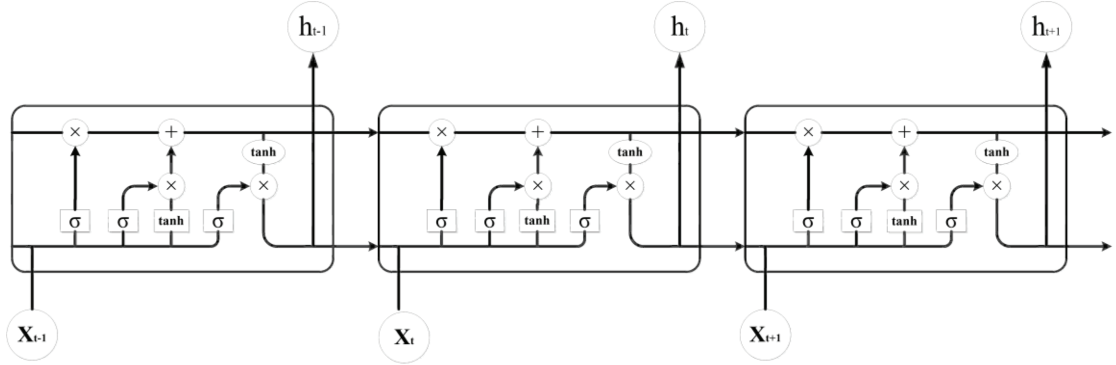

Fig. 2: LSTM network structure.

sequence data (Lipton et al. 2015).

The calculation equations of the recurrent neural network are shown in Equation (2) and Equation (3).

$$O_t = \mathrm{s}(V \cdot S_t) \qquad \ldots(2)$$

$$S_t = f(U \cdot X_t + W \cdot S_{t-1}) \qquad \ldots(3)$$

Where, *O* represents the value of the output layer, and *S* represents the value of the hidden layer. The weight matrix of the hidden layer to the output layer is represented by *V*. The weight matrix of input layer to the hidden layer is represented by *U*. *W* is the weight matrix with the last value as this time input of the hidden layer.

f () function is an activation function which can filter out unimportant information. The commonly-used activation functions are sigmoid function and tanh function, and the function expression is shown in Equation (4) and Equation (5).

$$f(z) = \frac{1}{1 + \exp(-z)} \qquad \ldots(4)$$

$$f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \qquad \ldots(5)$$

This study uses tanh as the activation function, and the LSTM network structure is shown in Fig. 2.

Where, x_t and h_t represent the inputs and outputs of the current cyclic neural network, respectively.

The calculation of the input gate during forwarding propagation is according to Equation (6); Equation (7) shows the calculation of the forgetting gate, and the output gate is based on Equation (8).

$$a_l^t = \sum_{i=1}^{I} w_{il} x_i^t + \sum_{h=1}^{H} w_{hl} b_h^{t-1} + \sum_{c=1}^{C} w_{cl} s_c^{t-1} \qquad \ldots(6)$$

$$a_{\mathrm{j}}^t = \sum_{i=1}^{I} w_{i\mathrm{j}} x_i^t + \sum_{h=1}^{H} w_{h\mathrm{j}} b_h^{t-1} + \sum_{c=1}^{C} w_{c\mathrm{j}} s_c^{t-1} \qquad \ldots(7)$$

$$a_w^t = \sum_{i=1}^{I} w_{iw} x_i^t + \sum_{h=1}^{H} w_{hw} b_h^{t-1} + \sum_{c=1}^{C} w_{cw} s_c^{t-1} \qquad \ldots(8)$$

Where, w_{ij} represents the connection weight from nerve cell *i* to nerve cell *j*. The input of nerve cells is represented by "a" and their output is indicated by "*b*". Subscripts *l*, j and w refer to the input gate, output gate, and forgetting gate respectively. Subscript *c* refers to the cell. The peephole weights from cell to the input gate, forgetting gate, and output gate are w_{cl}, $w_{c\mathrm{j}}$, and $w_{c\mathrm{w}}$ respectively. s_c refers to the status of cell *c*. *f* refers to the activation function of the control gate, and *h* is the output activation function of the cell.

During back propagation, suppose that:

$$\mathrm{e}_c^t = \frac{\partial l}{\partial b_c^t} \qquad \mathrm{e}_s^t = \frac{\partial l}{\partial s_c^t} \qquad \ldots(9)$$

Then, the calculation of the output gate is as shown in Equation (10) and the status is shown in Equation (11).

$$\mathrm{d}_w^t = f'(a_w^t) \sum_{c=1}^{C} h(s_c^t) \mathrm{e}_c^t \qquad \ldots(10)$$

$$\mathrm{e}_s^t = b_w^t h^{'}(s_c^t) \mathrm{e}_c^t + b_{\mathrm{j}}^{t+1} \mathrm{e}_s^{t+1} + w_{cl} \mathrm{d}_l^{t+1} + w_{c\mathrm{j}} \mathrm{d}_{\mathrm{j}}^{t+1} + w_{cw} \mathrm{d}_w^t \qquad \ldots(11)$$

The Adam algorithm was used for model optimization in this study. The squared-gradient normalized learning rate was used for gradient descent in the Adam algorithm (Kingma 2015). Firstly, a first-order momentum equivalent to the gradient was maintained. Then a momentum equivalent to the squared gradient was maintained. At last, the learning rate was normalized using the squared gradient *v,* as shown in Equation (12). The update amplitude is gradient *m* which is shown in Equation (13).

$$v_t = \mathrm{b} v_{t-1} + (1 - \mathrm{b}) \cdot \nabla Q(w)^2 \qquad \ldots(12)$$

$$m_t = \mathrm{a} m_{t-1} + (1 - \mathrm{a}) \cdot \nabla Q(w) \qquad \ldots(13)$$

Table 2: Basic Code Framework of LSTM.

Function	Code framework
Model-based category	nn.Module
Establish RNN layers	nn.LSTM (input_size, hidden_size, num_layers)
Establish linear layers	nn.Linear (hidden_size, output_size)
Loss function	torch.nn
Optimization function	torch.optim
Back propagation	optimizer.zero_grad loss.backward optimizer.step

Table 3: Parameter Setting.

Parameter	Parameter value
Training set	The data of 2015-2017
Test set	The data of 2018
Activation function of the hidden layer	Sigmoid
Batch size	3
Input dimension	2
Hidden layer dimension	4
Number of network layers	3
Activation function of the output layer	tanh
Learning rate	0.01

LSTM Algorithm Framework and Parameter Setting

Algorithm framework: The training algorithm of LSTM is a back-propagation algorithm. According to the above LSTM model algorithm, the steps are as follows.

Step 1: Calculate the output value of each nerve cell through forwarding propagation.

Step 2: Calculate the error value of each nerve cell through back-propagation.

Step 3: Calculate the weighted gradient.

In this paper, the Pytorch was applied for all experiments. The Pytorch is a Python-first deep learning framework which can realize acceleration on GPU, with the function of a dynamic neural network. The basic code framework of LSTM algorithm is shown in Table 2.

Parameter setting: We took the 2015-2017 data as a training set for predicting the data of 2018. For details of parameters are set in Table 3.

RESULTS AND DISCUSSION

Model Prediction Results

We predicted based on the above parameters and compared the prediction results of the five cities with the real values in 2018. Besides, we provided a decreasing figure of loss functions in the iteration. The results are shown in Fig. 3-7.

According to Figs. 3-7, the prediction results are very close to the real values and the loss ratio finally outputted by the loss function after 1000 times of iteration is close to 0. It can be seen from the trend of $PM_{2.5}$ concentration in 2018, the data (recorded or something else) for these five cities are the lowest in summer and the highest in winter.

Output Analysis

Hereinbefore, we have obtained a predicted model and outputted the predicted results. Table 4 shows the test results of the predicted values and the real ones.

The coefficient of association between the $PM_{2.5}$ means, standard deviations, the correlation coefficient between real values and predicted values in 2018 of the five cities are given in Table 4. The mean values of the real values and predicted values are approximate. According to the standard deviation, the $PM_{2.5}$ concentration data of Beijing, Chengdu and Wuhan are very discrete. It indicates that there is a large fluctuation

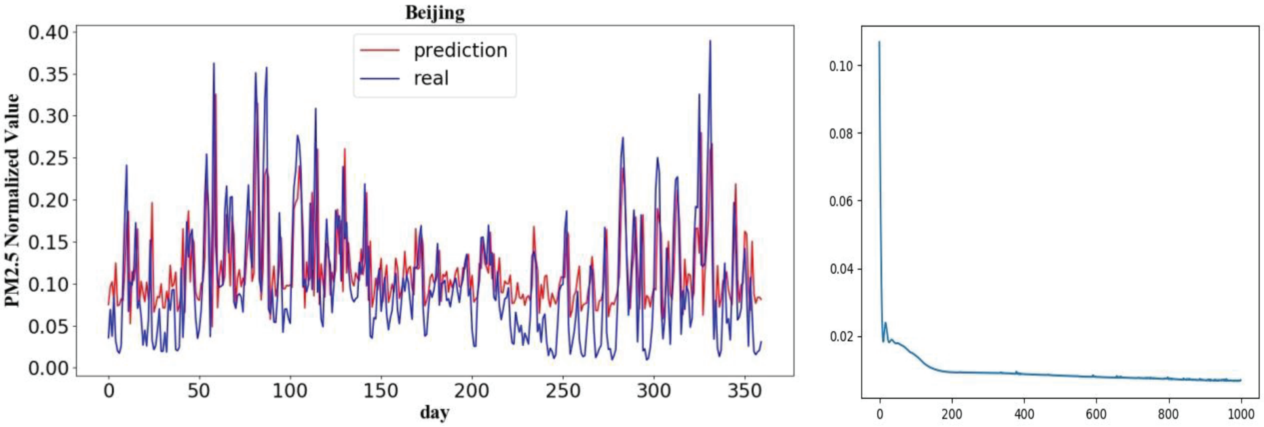

Fig. 3: Model prediction result of Beijing City. (a) Comparison between prediction values and real values (b) Loss rate during iteration.

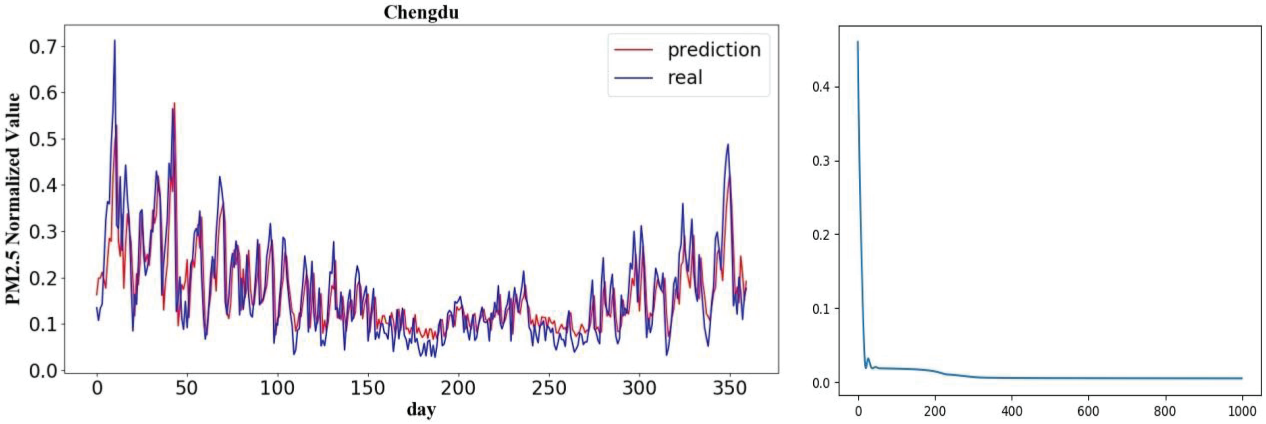

Fig. 4: Model prediction result of Chengdu City. (a) Comparison between prediction values and real values (b) Loss rate during iteration.

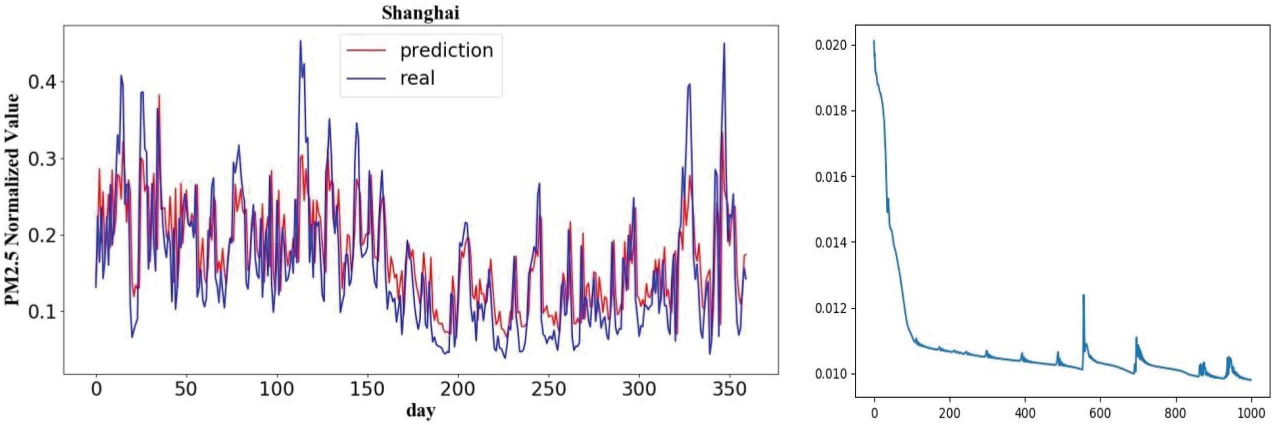

Fig. 5: Model prediction result of Shanghai City. (a) Comparison between prediction values and real values (b)Loss rate during iteration.

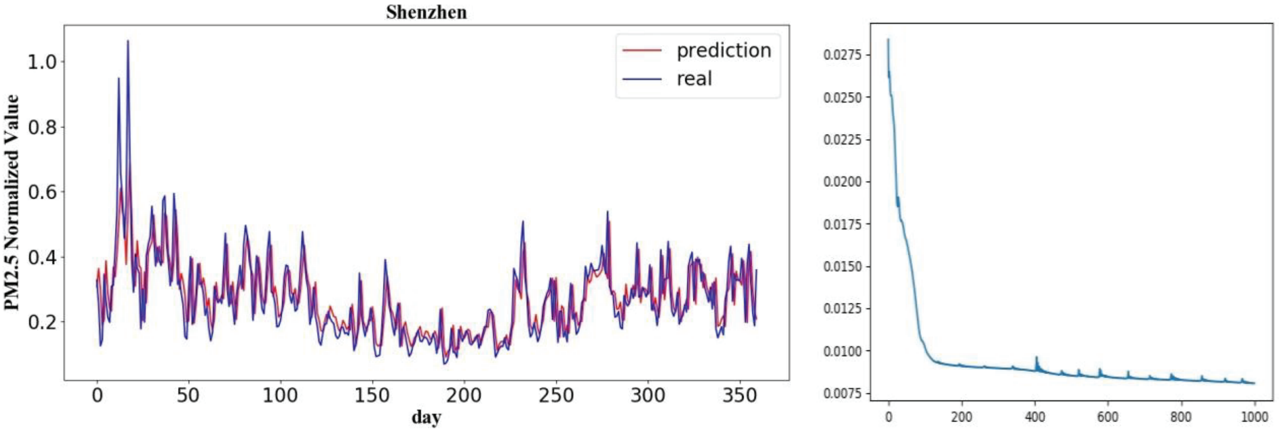

Fig. 6: Model prediction result of Shenzhen City. (a) Comparison between prediction values and real values (b) Loss rate during iteration.

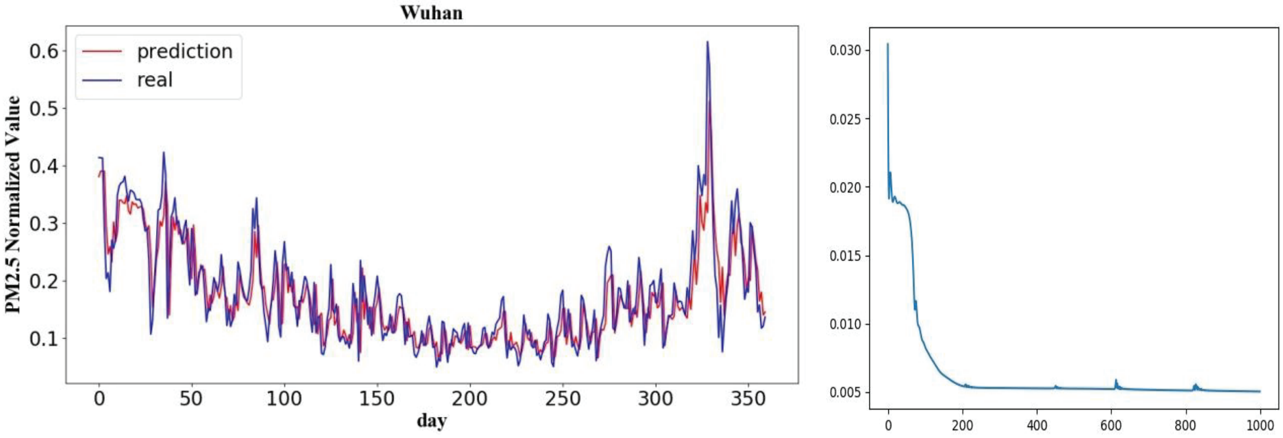

Fig. 7: Model prediction result of Wuhan City. (a) Comparison between prediction values and real values (b) Loss rate during iteration.

in the $PM_{2.5}$ concentration of these three cities. From the coefficient of association, except that the correlation coefficient of Beijing between the real values and predicted ones is 0.64118, the coefficients of the other four have exceeded 0.7, while that of Wuhan reaches 0.86724. This shows a relatively strong association. It also indicates that the trend of predicted values is the same as that of the real ones.

CONCLUSION

In this paper, by using the LSTM model, the $PM_{2.5}$ concentrations in five Chinese cities with GDP reaching 1000 billion Yuan were predicted based on the $PM_{2.5}$ data selected from 2015-2017 Based on the training results, the number of hidden neurons and neural network layers, the learning rate, and other parameters were adjusted. Then, the $PM_{2.5}$ concentration in 2018 was predicted after 1000 times of iteration and compared with the real $PM_{2.5}$ data collected in 2018. According to the comparison results, the real values and predicted values are approximately the same, and the coefficients of association between the real values and predicted ones of all cities are around 0.8. The predicted results have relatively high reliability and universality.

The LSTM model studied in this paper can be used to realize the good predicted of $PM_{2.5}$ concentration data with relatively high fluctuation. The model is applicable to cities with large air quality difference, especially the large cities with rapid industrial development. As an accurate method for predicting urban air quality, the LSTM model can help large cities to reduce the harm of $PM_{2.5}$ pollution to the human body by taking timely measures to predict and control $PM_{2.5}$ pollution.

ACKNOWLEDGEMENTS

This study was supported by the Fundamental Research Funds

Table 4: The test results of the predicted values and the real ones.

City	Real Value Predicted value	Mean value	Standard deviation	Correlation coefficient
Beijing	Real Value	45.43884	31.55345	0.64118
	Predicted value	48.81304	21.91014	
Chengdu	Real Value	47.06011	29.95005	0.80070
	Predicted value	48.50343	24.61025	
Shanghai	Real Value	31.99997	17.01369	0.72147
	Predicted value	33.00228	11.96767	
Shenzhen	Real Value	25.22512	11.71571	0.78225
	Predicted value	25.54528	9.60629	
Wuhan	Real Value	47.94295	24.91919	0.86724
	Predicted value	46.56945	21.82955	

for the Central Universities of China (No. KYTZ201661), China Postdoctoral Science Foundation (No. 2015M571782), and Jiangsu Agricultural Machinery Foundation (No. GXZ14002), Student Research Training Project of College of Engineering, Nanjing, Agricultural University (No. 1830C29). Thanks to Qingyue Open Environmental Data Centre (https://data.epmap.org) for support on Environmental data processing.

REFERENCES

An, F., Yuan, Y., Ma, X. and Shen, N. 2018. Tilapia growth prediction model based on Long Short-term Memory neural network. Journal of Southern Agriculture, 49(10): 2110-2116.

Bai, H., Shen, R., Shi, H. and Dong, Y. 2013. Forecasting model of air pollution index based on BP neural network. Environmental Science & Technology, 36(3): 186-189.

Christakos, G. and Serre, M.L. 2000. BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmospheric Environment, 34(20): 3393-3406.

Evans, J., van Donkelaar, A., Martin, R.V., Burnett, R., Rainham, D.G., Birkett, N.J. and Krewski, D. 2013. Estimates of global mortality attributable to particulate air pollution using satellite imagery. Environ. Res., 120: 33-42.

Fu, L., Yang, M. and Chen, Y. 2016. Factors influencing $PM_{2.5}$ and the governance strategies in Jiangsu, China. Nature Environment and Pollution Technology, 15(4): 1401-1408.

Graves, A. 2008. Supervised Sequence Labelling with Recurrent Neural Networks. Ph.D. Dissertation.

Guo, H., Wang, X., Wu, B. and Li, X. 2016. Cognizing population density demarcative Line (Hu Huanyong-Line) based on space technology. S&T and Society, 31(12): 1385-1394.

He, X., Lin, Z., Liu, H. and Qi, X. 2016. Analysis of the driving factors of $PM_{2.5}$ in Jiangsu province based on grey correlation model. Acta Geographica Sinica, 71(7): 1119-1129.

Kingma, D.P. and Ba, J.L. 2015. Adam: A method for stochastic optimization. International Conference on Learning Representations, 1-15.

Li, S. 2018. LSTM recurrent neural network short-term power load forecasting based on TensorFlow. Shanghai Energy Conservation, 12: 974-977.

Lipton, Z.C., Berkowitz, J. and Elkan, C. 2015. A critical review of recurrent neural networks for sequence learning. Computer Science, 1-38.

Liu, Q. 2016. Short term stock price forecasting based on fuzzy deep learning network algorithm. Harbin Institute of Technology.

Qiao, S., Sun, R. and Liu, J. 2017. Short-term traffic flow forecast based on deep learning. Journal of Qingdao University (Natural Science Edition), 30(4): 65-69.

Quan, B., Yang, B., Hu, K., Guo, C. and Li, Q. 2018. Prediction model of ship trajectory based on LSTM. Computer Science, 45(11): 126-131.

Sun, M. 2018. Prediction model of air quality index based on optimized deep belief network. China University of Geosciences (Beijing).

Wang, Q., Xia, S., Wan, Y. and Jin, L. 2009. A new idea for urban air pollution forecast. Environmental Science & Technology, 32(3): 195-198.

Yin, W., Zhang, D., Yan, J., Zhang, C., Li, Y. and Rui, X. 2015. Deep learning based air pollutant forecasting with big data. Chinese Journal of Environmental Management, 7(6): 46-52.

Zhou, X., Su, X. and Yuan, M. 2004. Forecast of air pollution index based on BP neural network. Journal of Harbin Institute of Technology, 36(5): 582-585.