Air Quality Prediction of Three Provinces in Central China Based on Hybrid K-Means-LSTM

Guoqu Deng, Hu Chen*, Siqi Wang

School of Management, Henan University of Science and Technology, Luoyang, China
**Corresponding Email: 200320190877@stu.haust.edu.cn*

Abstract

With the development of industrialization and urbanization, air pollution is becoming more and more serious. To protect people's health, reasonably predict air quality and provide suggestions for people's egress, this article constructs a short-term air quality prediction model based on K-Means-LSTM. The results show that the daily meteorological data of three provinces in Central China, Henan Province, Hubei Province and Hunan Province from 2017 to 2019 are selected, the daily average AQI is taken as the target variable, and the provincial capital city of each province from October to December 2019 is selected as the test data. The prediction accuracy of K-Means-LSTM model is better than LSTM, BPNN and XGBoost, indicating the practicability of the model proposed in this research.

Keywords: *Air quality prediction; Data clustering; K-Means-LSTM; Central China*

1. INTRODUCTION

With the development of industrialization and urbanization, air pollution is becoming more and more serious. To protect people's health, reasonably predict air quality and provide suggestions for people to go out for a long time, Air quality prediction has always been one of the most important research areas in the field of air pollution monitoring and control for current domestic and foreign scholars. Classified by existing research means, air quality prediction can be divided into three schemes: statistical prediction, numerical prediction and artificial intelligence algorithm prediction represented by machine learning [1].

Statistical prediction refers to the prediction by analyzing the change law of pollutants in the air, the commonly used models are the gray model and regression model [2, 3]. Numerical prediction is based on the law of atmospheric motion, combining the possible physical and chemical changes between pollutants in the air, the mathematical method is used to establish a pollutant diffusion model to predict the change of pollutant concentration of the future, and the mainstream models to include CQMA model and NAQPMS model [4]. Although mathematical statistics and numerical prediction models can achieve relatively good prediction results, however, the data of AQI is disordered and non-stationary, so these models are no longer applicable.

In recent years, with the rapid development of artificial intelligence, some scholars apply neural network to prediction in terms of air quality, the accuracy of prediction results is improved [5]. Some scholars have also constructed air quality prediction models based on recurrent neural network (RNN), and achieved higher prediction accuracy. However, there are some problems with RNN operation, such as gradient disappearance, short memory time series data and so on [6]. Long short-term memory (LSTM), neural network with continuous time feature extraction short term memory, shows better effect on AQI prediction with multi-dimensional variables. The commonly used LSTM algorithm is too time dependent and ignores the classification characteristics of the training data set in the process simulation and numerical prediction.

In summary, this research will build a short-term air quality prediction model based on K-Means-LSTM algorithm. Firstly, the K-Means algorithm is used to cluster the data, and the small sample data after clustering is input into LSTM as training data onto all sample prediction; Then, the collected prediction data is summarized, the data sorted by time is selected and reload into LSTM to obtain a new training model; Finally, some test data are selected to verify the excellence of the model, then it provides theoretical support for short-term air quality prediction [7-9].

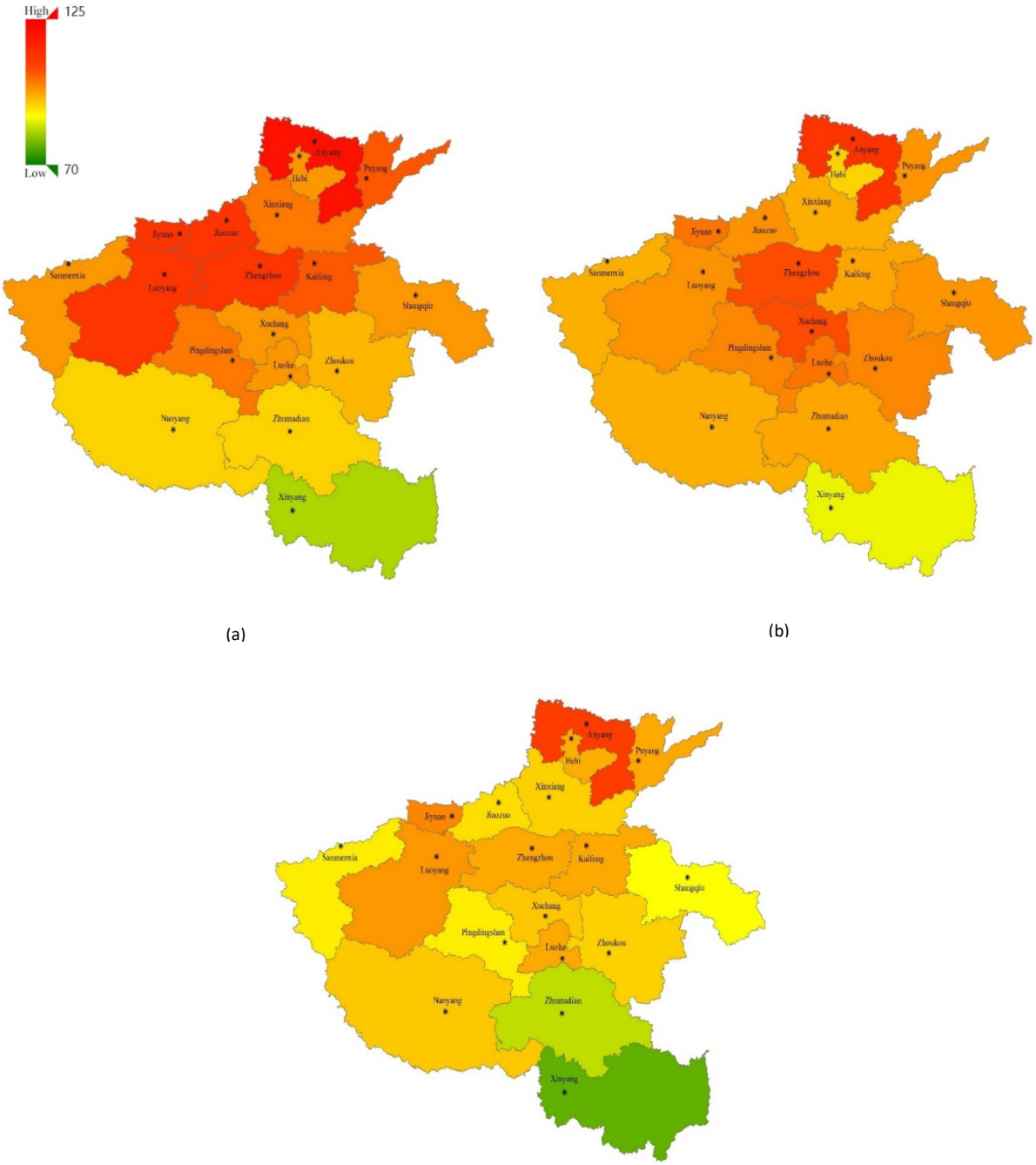
S. A. R. Khan et al. (Eds.): ICMSSE 2022, AHCS 12, pp. 135–142, 2023.
https://doi.org/10.2991/978-94-6463-056-5_19

2. DATA SOURCE

With the help of python and other relevant tools, this article captured the historical AQI and meteorological factors (temperature, wind direction, wind force and etc.) of cities in Henan Province, Hubei Province and Hunan Province from January 1, 2017 to December 31, 2019 published by the China Meteorological Administration, and drew the annual average AQI heat map, as shown in the Figure 1. At the same time, using the data collector to capture the daily concentrations of six pollutants in the three Province Capital cities of Zhengzhou, Wuhan and Changsha from January 1, 2017 to December 31, 2019 from the China Air Quality Monitoring and Analysis Platform. Finally, the sample data of AQI, meteorological factors and various pollutants of the three cities are integrated to carry out research.

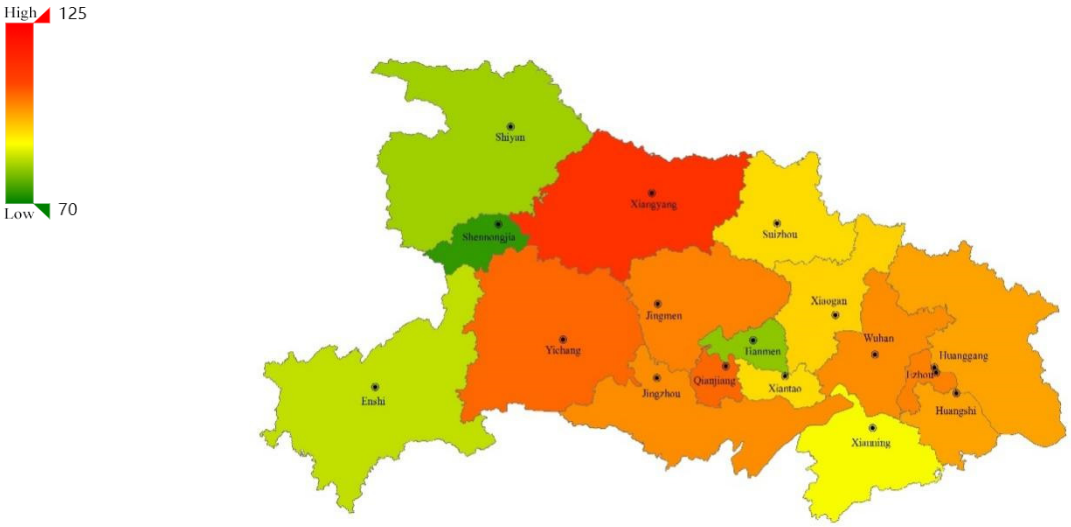
3. CONSTRUCTION OF PREDICTION MODEL

3.1. K-Means Clustering

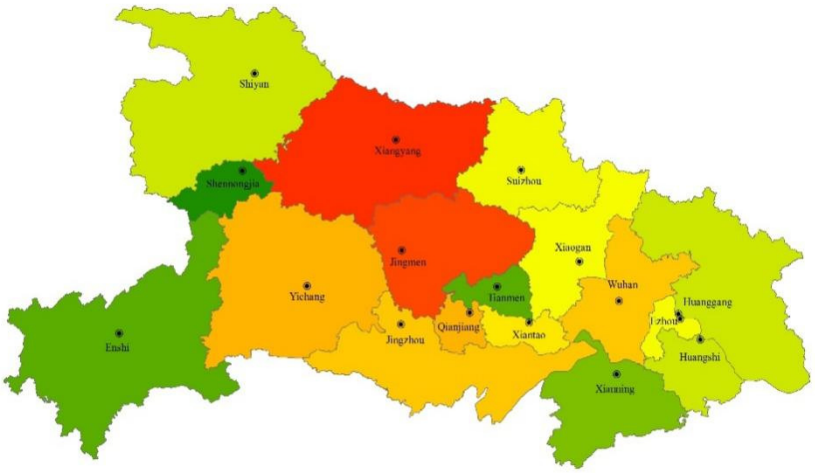


(a) (b)

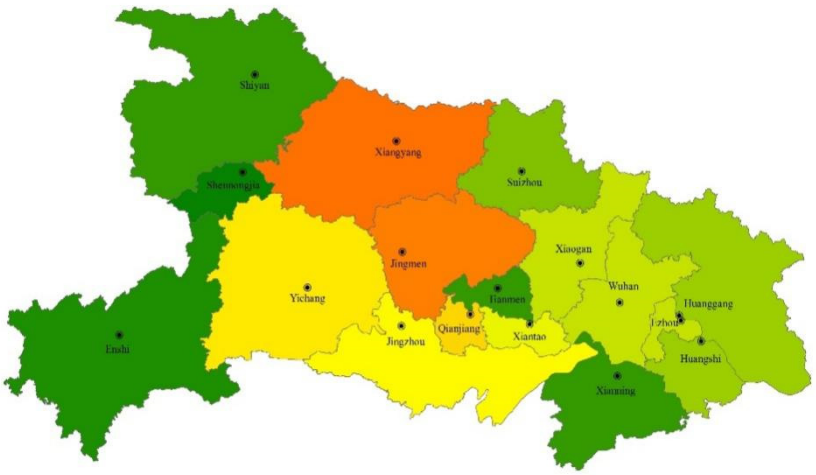
(c)


High
125
Low
70
Shiyan
Xiangyang
Suizhou
Shennongjia
Jingmen
Xiaogan
Wuhan
Yichang
Tianmen
Huanggang
Ezhou
Qianjiang
Xiantao
Jingzhou
Enshi
Huangshi
Xianning

(d)


Shiyan
Xiangyang
Suizhou
Shennongjia
Jingmen
Xiaogan
Wuhan
Yichang
Tianmen
Huanggang
Ezhou
Qianjiang
Xiantao
Jingzhou
Enshi
Huangshi
Xianning

(e)


Shiyan
Xiangyang
Suizhou
Shennongjia
Jingmen
Xiaogan
Wuhan
Yichang
Tianmen
Huanggang
Ezhou
Qianjiang
Xiantao
Jingzhou
Enshi
Huangshi
Xianning

(f)

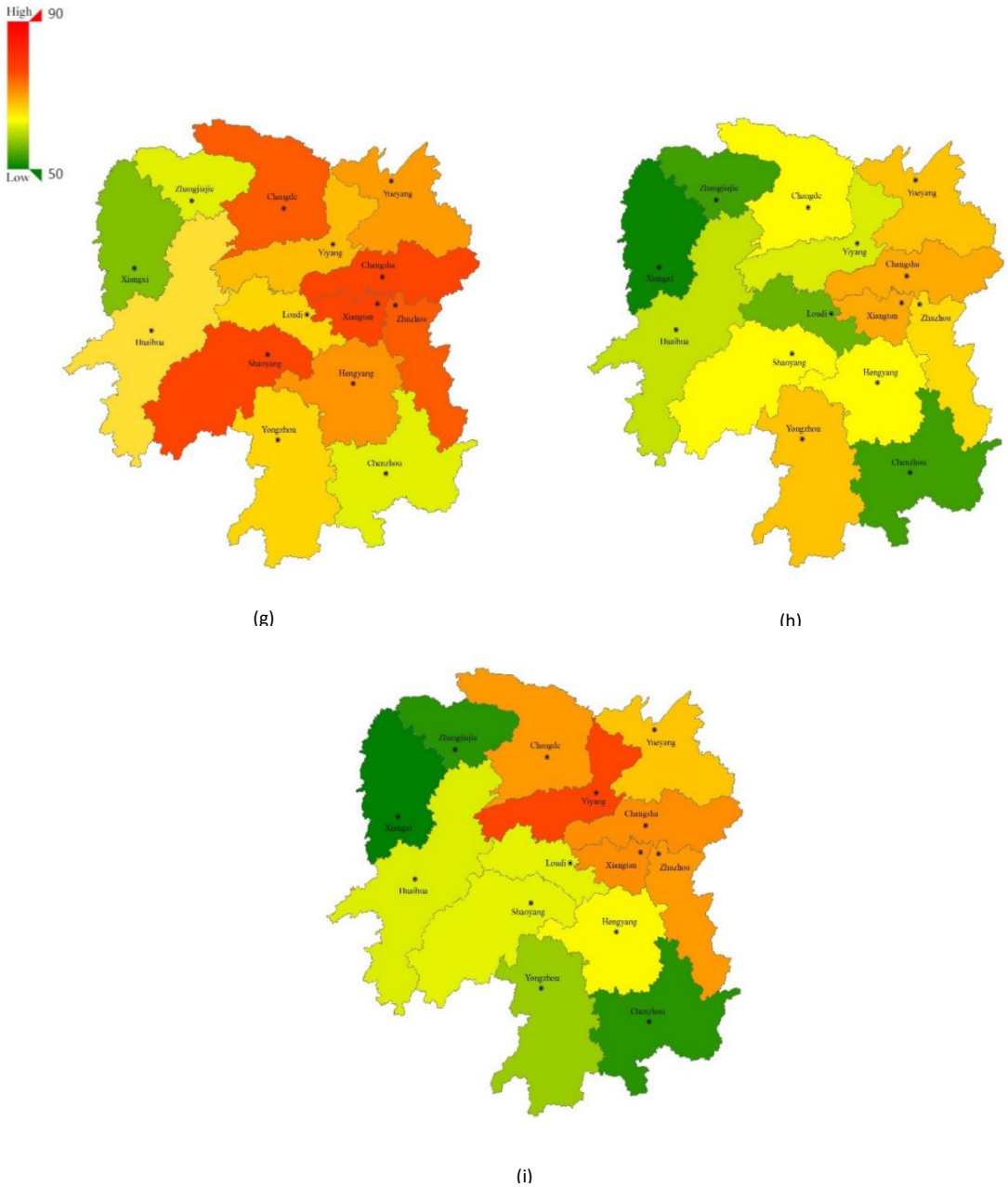


Figure 1. Annual average AQI heat map of each province from 2017 to 2019: (a), (b), (c): Henan. (d), (e), (f): HuBei. (g), (h), (i): Hunan.

To improve the regularity of data and shorten the prediction time, this study uses K-Means algorithm for data clustering. As a classical distance based clustering algorithm, K-means is widely used in various fields of prediction. The basic idea of the algorithm is as follows: Firstly, the algorithm needs to randomly points are assigned to the center closest to them to form the initial cluster; Finally, the average value of all points in each cluster are calculated. Take this average value as the new center point and repeat the process until the center of each cluster does not change [10-11]. The calculation formula is:

$$D(x, c_i) = \sqrt{\sum_{j=1}^{n} (x_j - c_{ij})^2} \quad (1)$$

Where x is the sample data, c_i is the i-th cluster center, n is the dimension of sample data, x_j is the j-th attribute of sample data, c_{ij} is the j-th attribute value and corresponding to the i-th cluster center.

The key step of K-means is to determine the number of clusters K. The selection criteria of this research are as follows: when the Euclidean distance from each point in the cluster center no longer changes significantly, calculate the sum of square error (SSE) of the corresponding points of the fitting data and the original data. The calculation formula is:

$$SSE = \sum_{i=1}^{k} \sum_{x \in S_i} \left(D(x, c_i)\right)^2 \quad (2)$$

Where D is the Euclidean distance between two points in the space, k is the number of clusters, and S_i represents the data set in the i-th cluster. When the SSE is closer to 0, the model selection is more and more better.

3.2.LSTM

The LSTM network structure includes memory unit cell, input door i_t and output door O_t, forgetting door f_t. The structure is shown in Figure 2. Part of the information of the unit cell state C_{t-1} is retained in the current cell state C_t, and the amount of retained information is determined by f_t. The specific learning process of LSTM is as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (3)$$

$$i_t = \sigma(W_i \cdot [h_t \cdot x_t] + b_i) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * C_t \quad (5)$$

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_0) \quad (6)$$

$$h_t = O_t * tanh(C_t) \quad (7)$$

C_t and h_t represent the activation vectors of each neuron cell and memory module, W and b represent weight matrix and offset vector, $*$ represents convolution. σ (·) indicates the activation function and tanh (·) denotes hyperbolic tangent function tanh ().

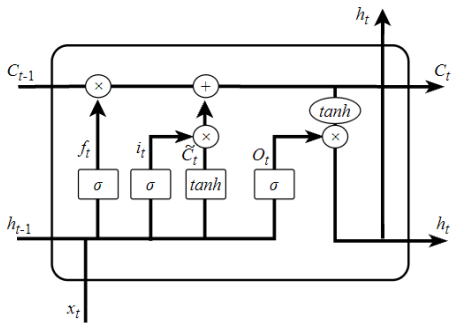


Figure 2. Schematic structure diagram of LSTM cell

In summary, the LSTM network can achieve better results for time series data. In this article, the K-Means-LSTM short-term air quality prediction model was constructed by combining K-Means and LSTM, thereby reducing the prediction effect of data differentiation on the model.

4. RESULT AND COMPARATIVE ANALYSIS

4.1.Evaluating Indicator And Analysis

This research chooses the root mean square error (RMSE) and mean absolute percentage error (MAPE) to test the prediction accuracy. The RMSE and MAPE can be expressed as:

$$RMSE = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (x_i^* - x_i)^2} \quad (8)$$

$$MAPE = \sum_{i=1}^{n} \left| \frac{x_i^* - x_i}{x_i} \right| \times \frac{100}{n} \quad (9)$$

where n represents the number of time instances, and x_i^* and x_i represent the real value and prediction data, respectively.

TABLE I. COMPARISON OF EVALUATION INDEXES OF EACH MODEL

City Name	Evaluation Indexes							
	K-Means-LSTM		*LSTM*		*BPNN*		*XGBoost*	
	RMSE	*MAPE (%)*	*RMSE*	*MAPE (%)*	*RMSE*	*MAPE (%)*	*RMSE*	*MAPE (%)*
Zhengzhou	5.23	9.31	7.18	15.24	13.19	18.96	10.07	14.78
Wuhan	4.95	8.26	8.74	17.36	10.78	19.33	16.24	30.52
Changsha	4.06	9.05	8.25	16.17	15.92	27.15	12.56	18.73

From the evaluation indexes of each model, it shows that the prediction effect of the K-Means-LSTM combined model is better than LSTM model without data clustering, BPNN model and XGBoost model. At the same time, the AQI of other cities in Henan Province, Hubei Province and Hunan Province are tested, and the results show that the average value of RMSE are lower than 7.63 and the average value of MAPE are lower than 15.54, which further verifies that the model proposed in this study is suitable for short-term air quality prediction in cities. The prediction results of each model are shown in Figure 3.

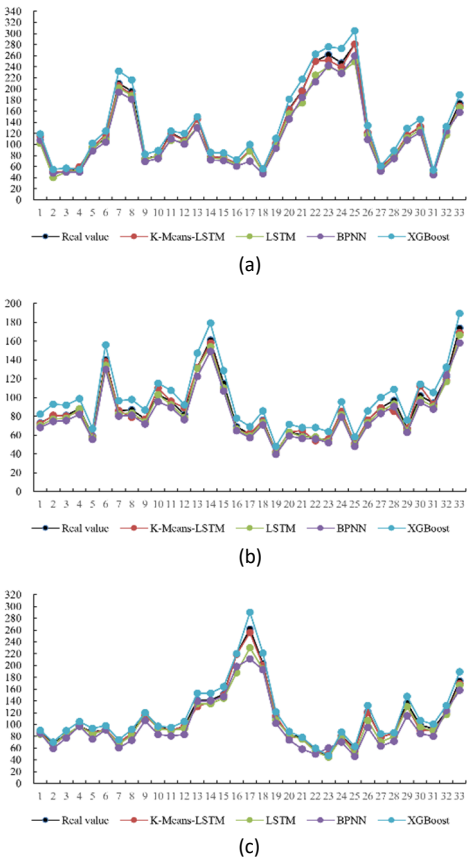


Figure 3. Predicted and real values of each model: (a): Zhengzhou. (b): Wuhan. (c): Changsha.

5. CONCLUSION

This article uses the K-Means clustering and LSTM to construct a short-term air quality prediction model the K-Means-LSTM, and takes three province capitals of Henan Province, Hubei Province and Hunan Province as examples to test the model. The test results show that the effect of the K-Means-LSTM is better than LSTM, BPNN and XGBoost model, and is more available for short-term air quality prediction. After clustering the data, the disadvantage of local over fitting of LSTM is avoided; Finally, using the above model to test other cities in each province, it is found that the effect of the K-Means-LSTM is still better than the other three models. Air quality is very important for human survival and development, therefore, it is particularly important to reasonably predict air quality and provide reference basis for government management and people's daily travel. Next, we will further explore the factors affecting AQI prediction and improve the optimization prediction algorithm to improve the prediction accuracy.

ACKNOWLEDGEMENT

This paper is greatly supported by National Social Science Fund Key Project (15AGL013); Henan Provincial Department of Science and Technology Risk Management Innovation and Public Policy Soft Science Research Base, Henan Social Science Planning Project (2019BJJ030); Henan Provincial Colleges and Universities Philosophy and Social Science Basic Research Major Project" Evaluation Research on Comprehensive Disaster Resilience Capacity of Chinese Communities" (2021JCZD04).

AUTHOR

Guoqu Deng, Hu Chen, Siqi Wang

Mailing address: School of Management, Henan University of Science and Technology, 263 Kaiyuan Avenue, Luolong District, Luoyang City, Henan Province, China

Postal Code: 471023

Telephone: 13014755062, 18437908006

REFERENCES

[1] Y. Li, Y. Bai and C, "Review on prediction model of air pollutant SO2," Sichuan Environment, vol. 35, no. 1, pp. 144-148, 2016.

[2] B. C. Zhang, "Prediction and analysis of air quality in Yinchuan City Based on residual modified GM (1,1) model," Green Technology, vol. 1, no. 12, pp. 118-122, 2019.

[3] Y. L. Zhang, Y. He, J. M. Zhu, "Research on PM2.5 prediction based on multiple linear regression model," Journal of Anhui University of Science and Technology, vol. 30, no. 3, pp. 92-97, 2016.

[4] R. Stern, P. Builtjes, M. Schaap, et al., "A model inter-com-parison study focussing on episodes with levated PM10 con- centrations," Atmospheric Environment, vol. 42, no. 19, pp. 4567-458, 2008.

[5] S. L. Zou, X. C. Ren, C. G. Wang, et al. "Effects of time accuracy and spatial information on neural network model prediction PM2.5 effect of concentration," Journal of Peking University (NATURAL SCIENCE EDITION), no. 3, pp. 417-426, 2020.

[6] Y. M. Zhu, A. L. Xu and Q. Sun, "New progress of air quality prediction method based on deep learning," China Environmental Monitoring, vol. 36, no. 3, pp. 10-18, 2020.

[7] X. L. Shi, L. Li, Q. H. Zhao, "Prediction of air quality index based on improved LSTM network," Statistics and Decision Making, vol. 37, no. 16, pp. 57-60, 2021.

[8] Y. M. Zhao, "LSTM algorithm and PM based on spatio-temporal correlation PM2.5 concentration prediction application," Computer Application and Software, vol. 38, no. 6, pp. 249-255, 323, 2021.

[9] Z. A. Ding, C. W. Le, L. L. Wu and M. L. Fu, "Fu Minglei. PM based on ceemd Pearson and depth LSTM hybrid model PM2.5 concentration prediction method," Computer Science, vol. 47, no. S1, pp. 444-449, 2020.

[10] K. J. Anil, "Data clustering: 50 years beyond k-means," Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010.

[11] C. H. Hung, H. M. Chiou and W. N. Yang, "Candidate groups search for K-harmonic means data lustering," Applied Mathematical Modelling, vol. 37, no. 24, pp. 10123-10128, 2013.

[12] Z. J. Yang, W. W. Yan, G. L. Wang, J. Y. Che, "Spatiotemporal prediction model of urban air quality based on big data," Control Engineering, vol. 27, no. 11, pp. 1859-1866, 2020.