Analysis Of Possible Influencing Factors and Forecasting the Concentration of PM2.5

Zijing Li*

School of Mathematics and Statistics, Wuhan University, Wuhan, China

* Corresponding Author Email: 2021302142022@whu.edu.cn

Abstract. Air quality is an important factor that affects the quality of people's lives. With the development of the times, people's health awareness is increasing. To truly enhance the well-being of all mankind, the author wants to figure out possible influencing factors of PM2.5 and forecast the concentration of it. The author gathers data from 2015 to 2023 and uses 4 different ways in analyzing influencing factors, then the author applies ARIMA model and LSTM model in forecasting. The author finds out that O3, precipitation, average 2-minute wind speed, and average relative humidity are nearly irrelevant to PM2.5 concentration; average air pressure and average temperature have no certainty effect on PM2.5 concentration; PM10, SO2, NO2, CO have a greater impact on PM2.5 concentration, and they are all positively correlated. As for forecasting models, LSTM model is better when performing long-step forecasts while ARIMA model is more accurate when performing short-step forecasts. These conclusions can be used by governments when forecasting relative parameters and they can take more accurate actions in dealing with PM2.5 pollution with the given influencing factors and their magnitudes.

Keywords: Correlation coefficients, Multiple linear regression model, ARIMA model, LSTM model.

1. Introduction

1.1. Research Background and Significance

Many scientists have studied relative issues about PM2.5 and its main resources, components and impacts on humans' health have already been thoroughly studied. However, there are still some issues that haven't been fully understood. For instance, the impact of PM2.5 on climate change is not certain, which needs to consider the radiative effect, cloud effect, precipitation effect and other aspects of PM2.5 and the control strategy and technology of PM2.5 need to be improved, which need to comprehensively consider the coordinated control of multiple pollutants, regional joint prevention and control, and synergistic effect of pollution reduction and carbon reduction. Based on these backgrounds, the author intends to find effective ways to improve air quality by studying the factors that influence PM2.5 concentration, and to find a more suitable and accurate prediction model for PM2.5 concentration by forecasting it, so as to enhance the precision of long-term and short-term prediction of PM2.5 for government and researchers.

1.2. Literature Review

In the article the researchers search for influencing factors and draw the conclusion that "The PD function of PM10 is approximately a step function, and the PD functions of other variables are approximately linear functions. The increase of these four variables causes different degrees of increase in PM2.5 concentration. Meanwhile, PM10 concentration also increases with NO2". This conclusion corresponds with the author's conclusion [1]. In the article, the writers draw the conclusion that it's precipitation and wind speed that lead to changes in PM2.5 concentration. The author reckons that the reason why the conclusion is different lies in the data used in those researches. The author uses a more comprehensive data while the essay's writers focuses on just one years' data in a particular region, so the climate factors is predominate in their research [2]. In the essay, the researchers use ARIMA model and AIC criterion, which is correspondent with the author's basic methods [3].

In the essay, the researchers draw the conclusion that in winter, PM2.5 shows a linear positive correlation with SO2 and NO2 respectively, and the concentration of PM2.5 is the highest in a weak low-temperature environment. This conclusion is roughly correspond with the author's conclusion, but due to regional difference, the conclusions are still slightly different [4].

1.3. Research Content and Framework

The research includes two sections. In section 1, the author takes a deep look at analyzing possible factors that influences the concentration of PM2.5 and section 1 is divided into 3 parts: methods, results and discussion. In section 2, the author uses ARIMA model and LSTM model in forecasting the concentration of PM2.5 and section 2 is also divided into 3 parts mentioned above. The author finds out the influencing factors of PM2.5 concentration, makes discussions about such results, compares the accuracy of different models in different predictions and analyzes the causes from the perspective of mathematical model principles. At the end, the author presents an conclusion based on this research.

2. Analysis of Possible Influencing Factors of PM2.5

The author gathers data and makes analyses based on this data set. Since there are some outliers in this data set, the author does data pre-processing first and eliminates the outliers. Then the author uses multiple linear regression model in analyzing the influencing factors of PM2.5. The author applies 4 different methods in analyzing this question: using Pearson correlation coefficient, Spearman correlation coefficient, multiple linear regression model and included angle cosine method in analyzing.

2.1. Pearson Correlation Coefficient

The Pearson correlation coefficient can be calculated by the following formula [5]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

2.2. Spearman Correlation Coefficient

The Spearman correlation coefficient can be calculated by the following formula:

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(R(x_i) - \overline{R(x)}) \cdot (R(y_i) - \overline{R(y)})}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})^2)(\frac{1}{n}\sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2)}} \tag{2}$$

2.3. Multiple Linear Regression Model

First, determine the target variable as the concentration of PM2.5, and the other variables are PM10, O3, SO2, NO2, CO, precipitation, average pressure, wind speed, temperature, and relative humidity, respectively. Use (a1, a2, …, an) to represent the coefficients of each factor in the target formula.

Second, normalize the variables of each factor to achieve the same dimension, calculate (a1, a2, …, an), and express the target variable as a linear expression determined by various variables [6]:

$$y = a_1 x_1 + a_2 x_2 + ... + a_n x_n + \varepsilon \tag{3}$$

which is the multivariate linear fitting line of y, and the absolute value of the coefficient reflects the degree of influence. After standardization, the degree of influence can be intuitively seen through the size of each factor's weight.

2.4. Using Included Angle Cosine in Analyzing

The author applies geometry methods and uses the cosine of the angle between 2 vectors to measure the size of the difference between 2 individuals [7].

2.5. Data visualization

Plot the curves of the four methods and the actual data, and visually assess the goodness of fit of the models from the graph, and provide a table of the relationship between the factors and the PM2.5 concentration for each research method.

3. Analysis of Correlation Coefficient Results

The author finds out the results by programming and the results are show in Table 1:

Table 1. The degree of influence of other factors on PM2.5 concentration

	Pearson	Spearman	Included angle cosine method	multiple linear regression model
PM10	0.8470	0.8813	0.8473	0.4198
O_3	-0.2462	-0.1994	-0.2463	0.0499
SO_2	0.5206	0.5735	0.5208	0.0009
NO_2	0.6096	0.6917	0.6098	0.0895
CO	0.7467	0.7013	0.7470	0.1716
Amount of precipitation	0.0096	-0.2316	0.0096	0.0242
Average value of air pressure	0.0474	0.4950	0.0474	0.0754
Average 2-minute wind speed	0.0445	-0.2864	0.0446	0.0115
Average temperature	0.0471	-0.5182	0.0472	0.0956
Average relative humidity	0.0472	-0.0198	0.0473	0.0616

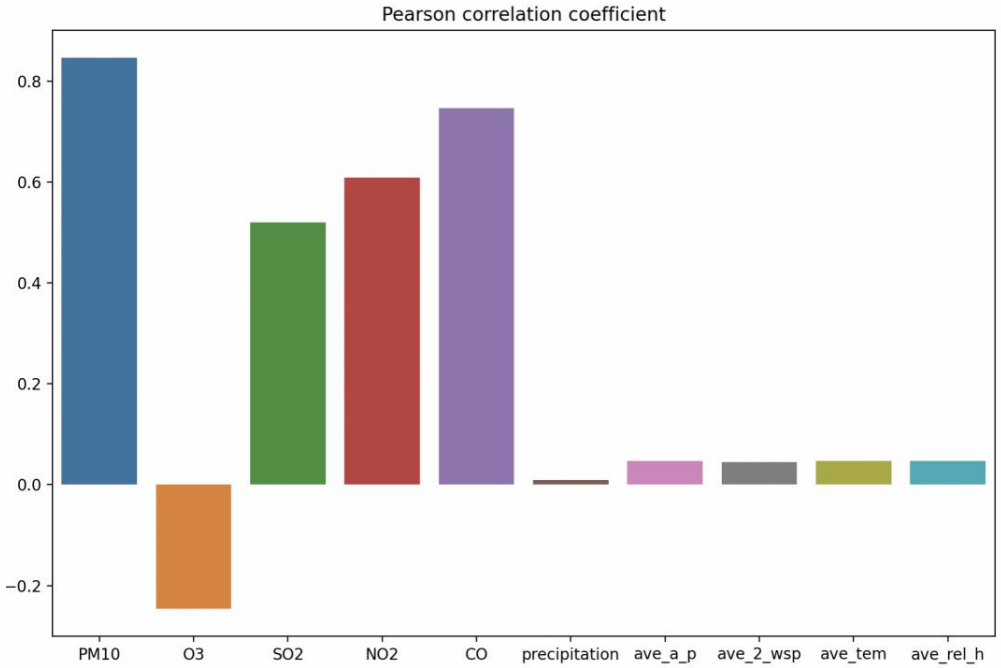


Fig. 1 The graph of Pearson correlation coefficient

As can be seen from the Figure 1, PM10, CO, NO_2, SO_2 and O_3 have great effect on the concentration of PM2.5 while precipitation, average value of air pressure, average 2-minute wind

speed, average temperature and average relative humidity have little impact on the concentration of PM2.5.

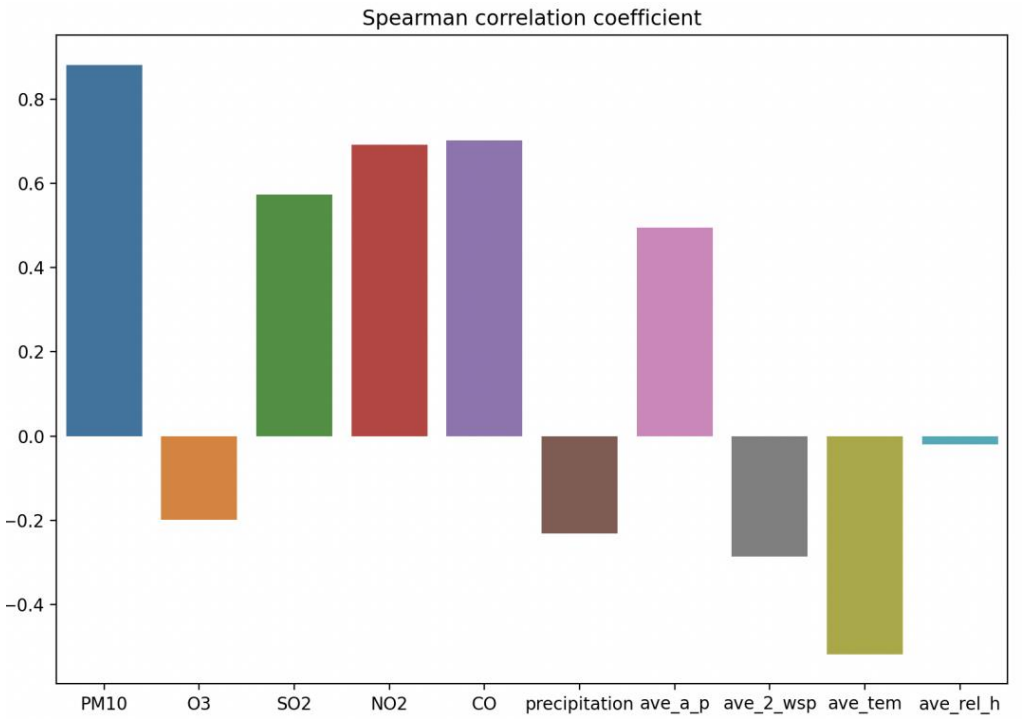


Fig. 2 The graph of Spearman correlation coefficient

As can be seen from the Figure 2, PM10, CO, NO_2, SO_2, average value of pressure and average temperature have great effect on the concentration of PM2.5 while precipitation, average 2-minute wind speed, O_3 and average relative humidity have little impact on the concentration of PM2.5.

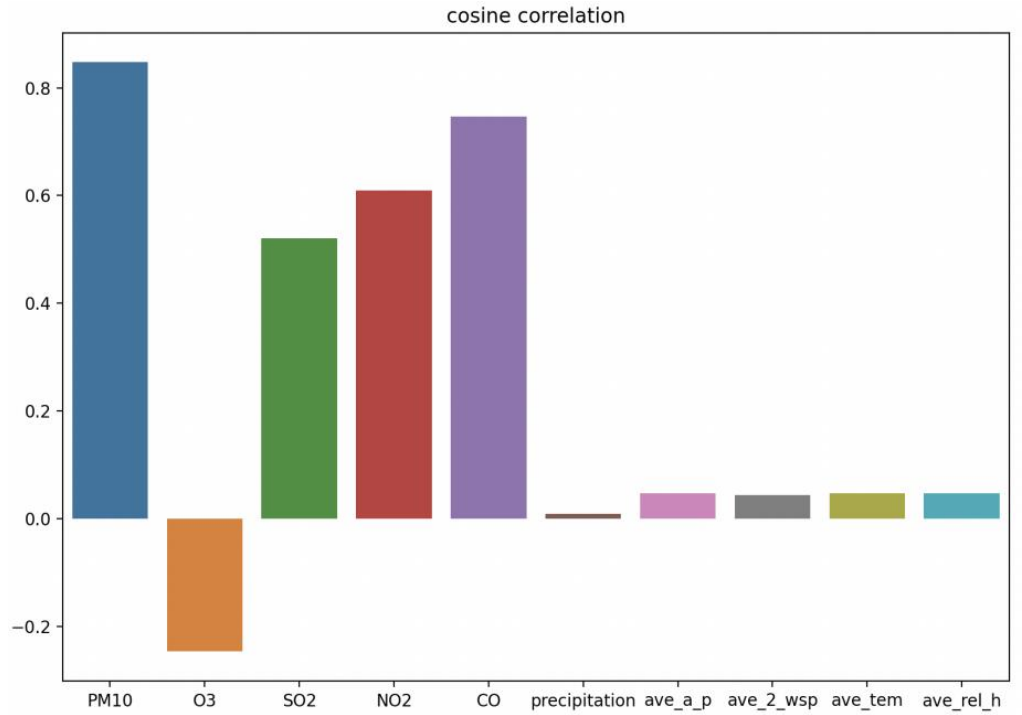


Fig. 3 The graph of included angle cosine method

As can be seen from the Figure 3, PM10, CO, NO_2, SO_2 and O_3 have great effect on the concentration of PM2.5 while precipitation, average value of air pressure, average 2-minute wind speed, average temperature and average relative humidity have little impact on the concentration of PM2.5.

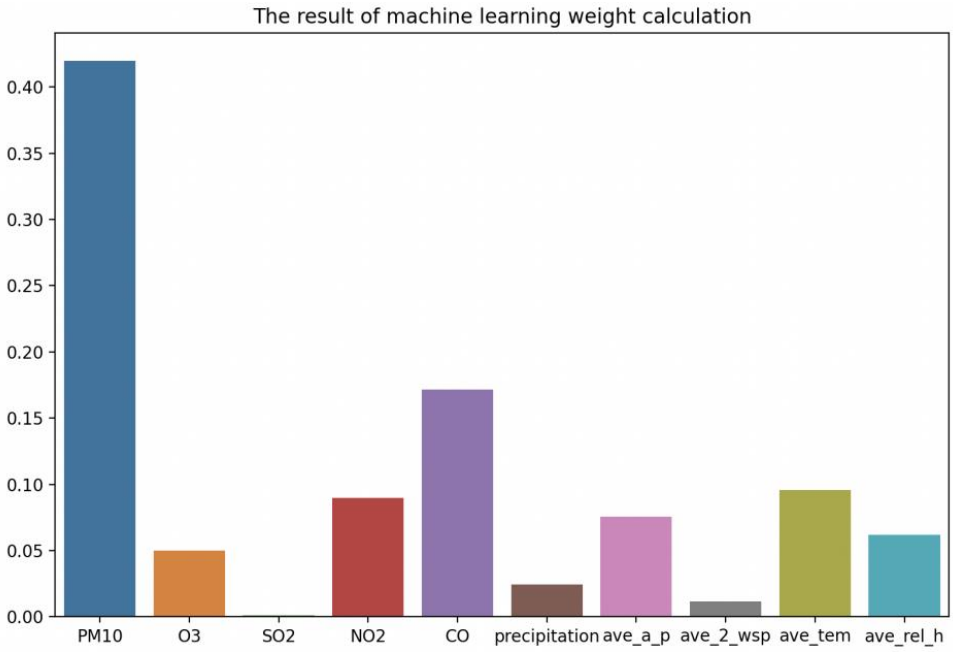


Fig. 4 The graph of multivariate linear regression coefficient weight graph (after standardization)

As can be seen from the Figure 4, PM10, CO, NO_2, O_3, average temperature, average value of air pressure and average relative humidity have great effect on the concentration of PM2.5 while precipitation, average 2-minute wind speed, and SO2 have little impact on the concentration of PM2.5.

The result of this question: Combining hypothesis 2, the conclusion is drawn: O3, precipitation, average 2-minute wind speed, average relative humidity are all irrelevant to PM2.5 concentration; average air pressure and average temperature have no certainty effect on PM2.5 concentration; PM10, SO_2, NO_2, CO have a greater impact on PM2.5 concentration, and they are all positively correlated. Among them, PM10 has the greatest positive correlation relation with PM2.5 concentration, the author thinks the reasons might be: since PM10 and PM2.5 both come from motor vehicles driving on unpaved, cemented roads, the crushing and grinding process of materials, dust raised by wind, etc, high concentration of PM10 indicates more possible resources of these particles, thus PM2.5 concentration is higher.

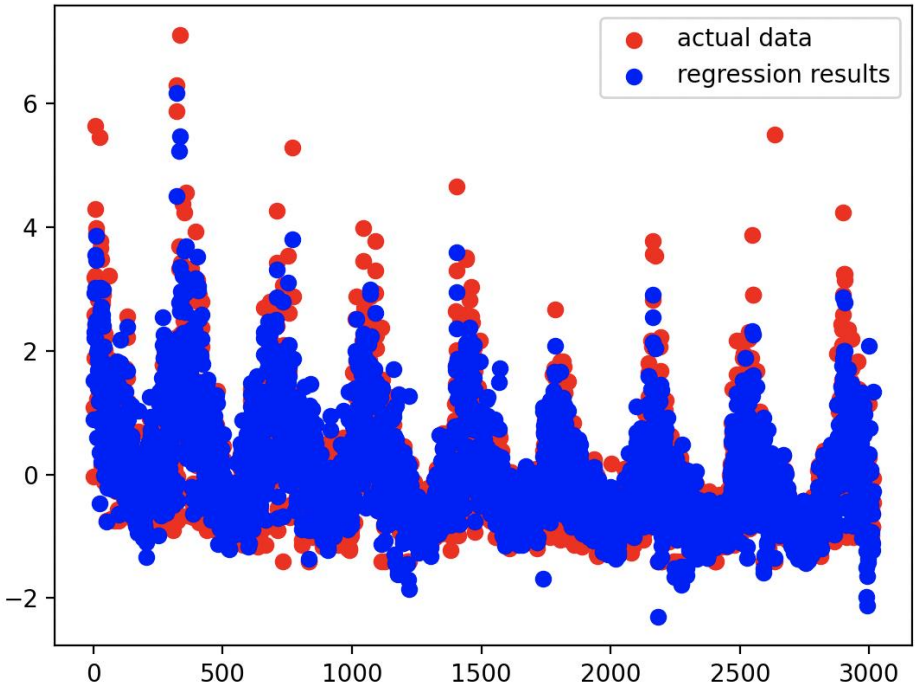


Fig. 5 Multivariate linear regression fitting results

The figure 5 shows the comparison of the results obtained by the multivariate linear regression model and the actual data, which indicates that the model has a good fitting effect. In addition, the coefficients of the regression equation are tested for significance, and the significance test levels of all coefficients are less than 0.05, indicating that the coefficient test is passed; in the collinearity diagnosis, the tolerance of the coefficients are all greater than 0.1 and the variance inflation factor (VIF) values are all less than 10, indicating that there is no multicollinearity phenomenon, making the data more convincing.

The analyses of the results obtained by four methods is explained by the author. Comprehensive analyses of the degree of influencing factors: All four calculation methods have reached the conclusion that PM10 has the greatest impact on PM2.5, and the ranking of other factors is slightly different, but they all conform to the argument that "the influence of air material composition on PM2.5 is higher than that of climate". The author searches for reasons of this conclusion. In terms of emission sources, PM2.5 mainly originates from the combustion of fossil fuels, industrial production, agricultural activities, biomass burning, transportation and other human activities, as well as natural phenomena such as volcanic eruptions, dust storms, forest fires, etc. Different emission sources produce different PM2.5 components and characteristics, affecting their reactions and deposition processes in the atmosphere [8]. In terms of chemical reactions, some gaseous substances in the air (such as SO2, NOx, VOCs, etc.) form secondary particles (such as sulfates, nitrates, organic carbon, etc.) under photochemical action, increasing the concentration of PM2.5. At the same time, various components of PM2.5 also undergo complex physical and chemical reactions in the air, changing their properties and effects[8]. Meteorological factors are greatly influenced by time and location, and even in the same region they have large fluctuations, so the influence of meteorological factors on PM2.5 concentration does not show a clear correlation like air material composition.

The author gives explanation of different mathematical methods: From the results, it can be seen that the results obtained by using Pearson correlation coefficient are exactly the same as those obtained by using cosine method, and the results obtained by using multivariate linear regression method and Spearman coefficient are slightly different from the previous two results. By analyzing the principles of these four methods, it is known that the principle of Pearson correlation coefficient is to measure whether two data sets are on a line, and describe the linear relationship between interval variables. This is essentially the same as the cosine algorithm, which can explain the high similarity of the results obtained by both. Spearman correlation coefficient is for data that do not follow normal distribution, raw data, rank data, one-sided open data, and data with unknown population distribution type that don't conform to use moment correlation coefficient to describe correlation. It essentially depicts the monotonic relationship between the two, which is different from the previous two methods in principle; multivariate linear regression method is to estimate the regression coefficients and intercepts by minimizing the residual sum of squares, so it also has some differences from the above methods. But combining with the actual situation to test the data obtained by these methods, the overall fit degree is high.

4. Building and Evaluating a Multi-step Prediction Model

4.1. Method

Since it's a time series model and the data doesn't have a distinct trend, the author considers applying ARIMA model and LSTM model in forecasting and analyzing. Firstly, the author chooses a data set partitioning method that has high forecasting accuracy. There are many methods, but the forecasting effects are different. The author evaluates the effects theoretically and comprehensively before determining the final partitioning scheme. Secondly, the author builds an ARIMA model for the determined training set and test set. This model requires pre-steps such as stationarity test, determining the order of difference, white noise test, and determining. The illustration of ARIMA model: ARIMA model is a combination of autoregressive moving average (ARMA) model and difference model. ARMA model is composed of autoregressive (AR) model and moving average

(MA) model. ARIMA(p,d,q) model is a model that transforms non-stationary time series into stationary time series by using difference method, and then regresses the dependent variable on its lagged values and the current and lagged values of random error terms. A model with the following structure is called an autoregressive moving average model, abbreviated as ARMA(p,q) model:

$$\Phi(B)(x_t - \mu) = \Theta(B)\varepsilon_t \tag{4}$$

where $\{ x_t \}$ is a stationary series with mean $E(x_t) = \mu$, B is a linear backward shift operator,

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p \tag{5}$$

$$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \tag{6}$$

$\{ \varepsilon_t \}$ is a stationary white noise with zero mean and variance σ_ε^2. In particular, if μ is unknown, it can be estimated by the mean value of x_t, $\bar{x}$.

Usually, the obtained series is non-stationary and needs to be differenced. If there exists a positive integer d, such that $\nabla^d x_t = w_t$, where $\{ w_t \}$ is an ARMA(p,q) series, then $\{ x_t \}$ is called an ARIMA(p,d,q) series. To elaborate, it satisfies [3,7]:

$$\begin{cases} \Phi(B)(\nabla^d x_t - \mu) = \Theta(B)\varepsilon_t, \quad t > d \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_s \varepsilon_t) = 0, \forall s < t \end{cases} \tag{7}$$

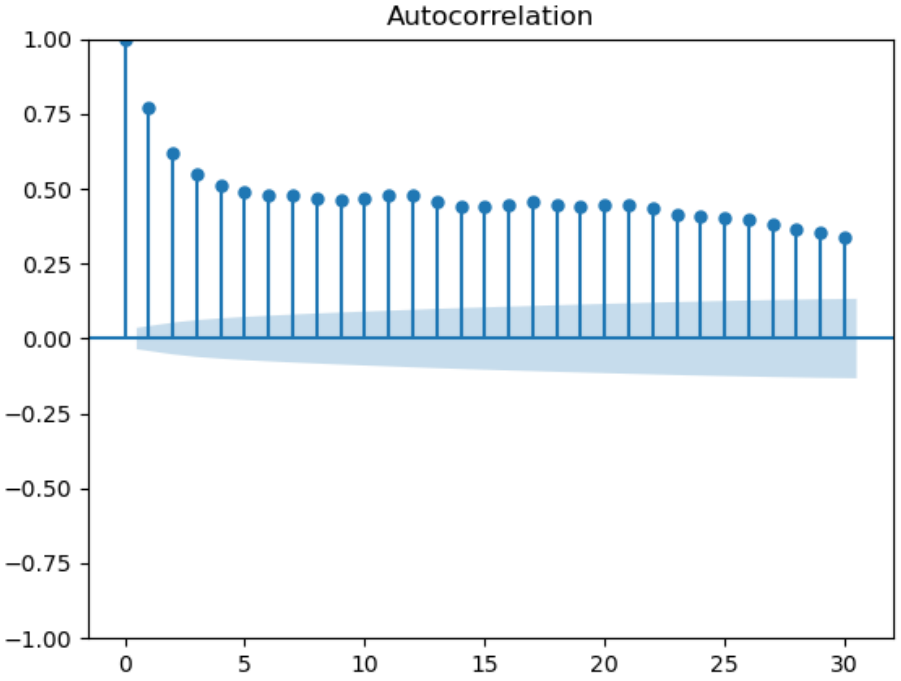


Fig. 6 Autocorrelation plot of PM2.5 concentration series (standard lag values)

As can be seen, it's neither a white noise nor stationary, so the author applies first difference to obtain a stationary time series.

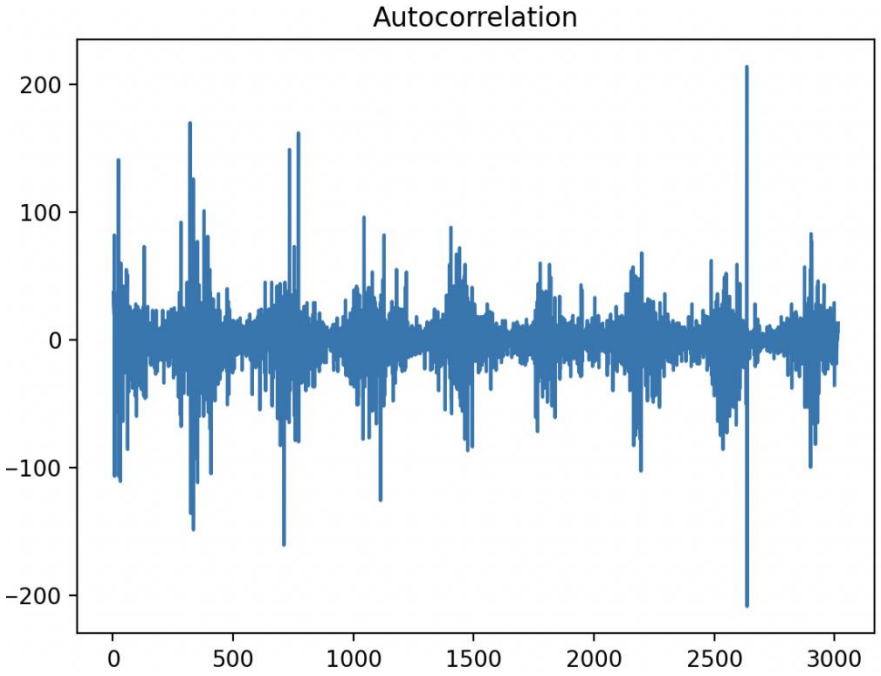


Fig. 7 First difference PM2.5 concentration time series plot

As shown in the figure 6 and 7, the time series plot of the new series obtained after the first difference fluctuates up and down around the mean of the difference data, which can be regarded as a stationary series. Further unit root test shows that the series does not have a unit root, and is a significantly stationary series. Then, the white noise test is performed on this series, and it is found that the series is a non-random series, which has predictive significance. Therefore, this model can determine d=1. The author then plots PACF and ACF graph and applies AIC criterion to determine the values of p and q.

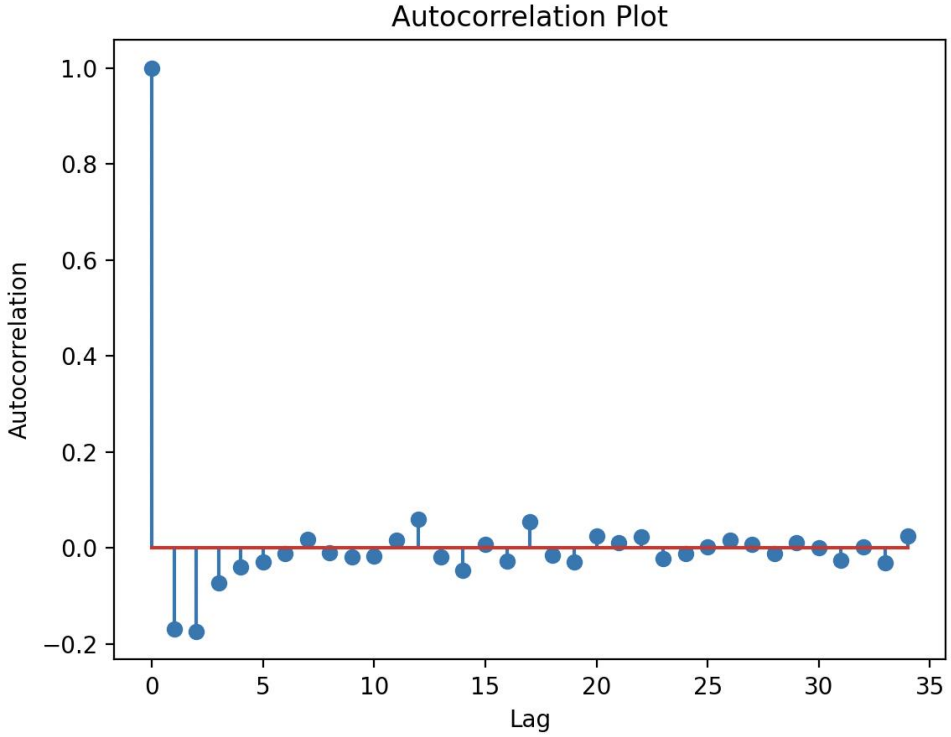


Fig. 8 ACF plot of first difference PM2.5 concentration

According to the ACF plot (figure 8), the author determines that p=2 or 0, to determine the specific value, the author applies AIC criterion and finds out that p=2 is the optimal solution [3].

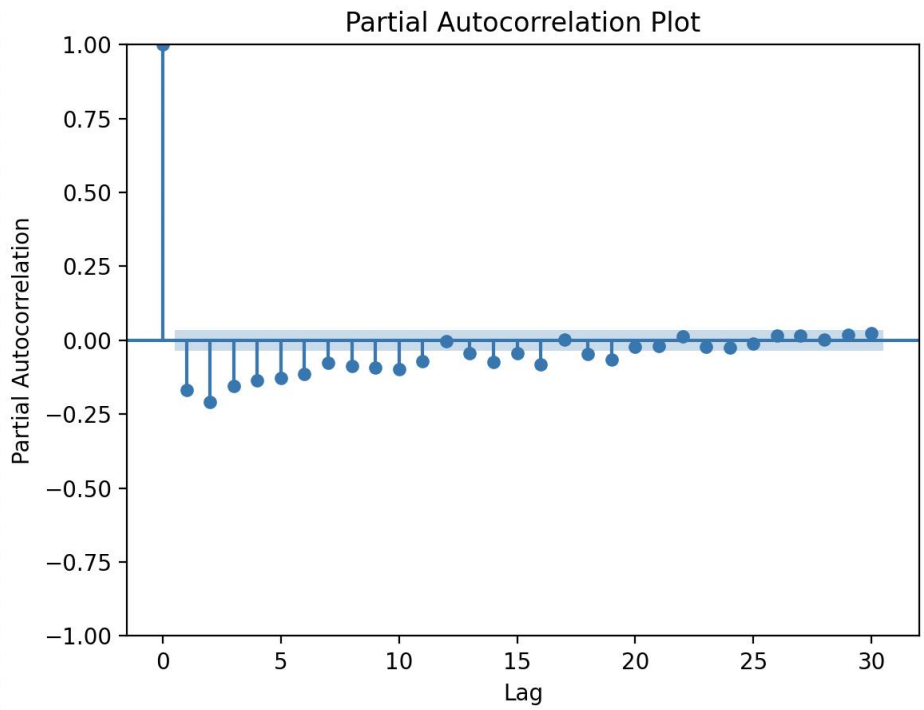


Fig. 9 PACF plot of first difference PM2.5 concentration

According to the figure 9 and AIC criterion, the author determines that q=2. The author then applies LSTM model in forecasting and makes comparison between the 2 models. LSTM model, short for Long Short-Term Memory Network, is a special type of Recurrent Neural Network, which can handle long time series data, and solve the problem of gradient vanishing or exploding in general Recurrent Neural Networks[9].

4.2. Results

The author performs 3-step、5-step、7-step、12-step forecasts using ARIMA model and calculates the value of RMSE, The results are as follows:

Table 2. RMSE value of multi-step forecasts (ARIMA model)

Length of forecasting	3-step forecast	5-step forecast	7-step forecast	12-step forecast
RMSE	3.59	6.17	8.08	8.32

As can be seen from these data, the value of RMSE is increasing with the length of steps, indicating that ARIMA model is accurate when being used in short-step forecasts, but errors start to accumulate as the length gets longer (in Table 2).

Table 3. The 12-step forecast results (ARIMA model)

date	2023.4.30	2023.5.1	2023.5.2	2023.5.3	2023.5.4	2023.5.5
PM2.5	40.6229	37.3063	34.9748	33.4484	32.5247	32.0199
date	2023.5.6	2023.5.7	2023.5.8	2023.5.9	2023.5.10	2023.5.11
PM2.5	31.7863	31.7146	31.7296	31.7832	31.8471	31.9065

Table 4. The results of 12-step forecast using LSTM model

date	2023.4.30	2023.5.1	2023.5.2	2023.5.3	2023.5.4	2023.5.5
PM2.5	30.1544	38.9382	40.8642	30.6134	28.3154	23.1878
date	2023.5.6	2023.5.7	2023.5.8	2023.5.9	2023.5.10	2023.5.11
PM2.5	25.4886	24.2976	23.3132	20.1815	22.1065	16.8032

The results of Table 2 show that the RMSE value is small, indicating that the model has a good forecasting effect; the smaller the step size, the smaller the RMSE value, which indicates that the ARIMA model has a better fitting effect in short-term forecast. The following will discuss this based on the principle of the model.

From the principle of ARIMA, when the time series data doesn't have an obvious upward or downward trend, the ARIMA model has a strong forecasting ability and a small forecasting error; when the time series data has a strong trend, it may cause the ARIMA model to have a large forecasting bias. The data doesn't have obvious tendency and regularity, so it conforms to the result that the model has a good forecasting ability. The ARIMA model assumes that the time series data is stationary in a statistical sense. The author makes the time series data stationary by using the first difference, so the ARIMA model fitting result has a relatively small and its forecasting effect is good. In the ARIMA model, the prediction result is affected by historical data, and historical errors will be magnified in the iteration process, so the longer the step size, the more times the error is iterated and magnified, and thus the lower forecasting accuracy (in Table 3). This explains why the ARIMA model has a good forecasting effect when the step size is short. The author checks the accuracy of 2 models and finds out that as for multi-step forecasts, LSTM model has better forecasting effect, the author thinks the reasons are: Compared with ARIMA model, LSTM model has more parameters and flexibility, and can adapt to different types and scales of time series data (in Table 4). LSTM model can also improve forecasting accuracy and robustness by using ensemble learning methods. Compared with LSTM model, ARIMA model has fewer parameters and restrictions, and needs more pre-processing and diagnosis for time series data. ARIMA model may also be less sensitive to the nonlinear relationship and outliers in time series data [10].

5. Conclusion

The author analyzes the factors affecting PM2.5 concentration and draws the conclusion that "the influence of gas molecules in the air is greater than that of climate", which is not exactly the same as some literature conclusions. The author analyzes that the reasons for this difference may include different data length selection, different research locations and time spans, etc. In addition, the author evaluates the four methods used in this chapter and explained the reasons for different conclusions from different methods from the perspective of mathematical model principles. The author uses ARIMA model and LSTM model to predict PM2.5 concentration, adopts multi-step prediction method and calculates the mean square error of ARIMA model prediction results, and draws the conclusion that "ARIMA model has high accuracy in short-step prediction, but not enough accuracy in long-step prediction". Using LSTM multi-step prediction model, it's found that the model could perform better multi-step prediction with longer step length. The author compares the two models and draws the following conclusions:

LSTM models can handle non-stationary and nonlinear data, while ARIMA models need to difference and transform data to meet the assumptions of stationarity and linearity. LSTM models can automatically extract data features, while ARIMA models need to manually select appropriate parameters (p,d,q) and seasonal factors. LSTM models can more flexibly handle multivariate and high-dimensional data, while ARIMA models are usually only suitable for univariate and low-dimensional data. LSTM models can better adapt to data changes and outliers, while ARIMA models may be affected by data noise and outliers. Of course, this does not mean that LSTM models are always better than ARIMA models. In some cases, ARIMA models may perform better. For example, when the data fluctuation is small, or has obvious periodicity and trend, ARIMA models may be more accurate and stable. While LSTM models may overfit or underfit, or require more computational resources and time. The author believes that this study has certain application value. First of all, from the perspective of the influencing factors of PM2.5 concentration, the author believes that local governments can take the prevention and control of related gas molecules as the primary goal, and take appropriate climate governance measures in different quarters and weather environments to

reduce the PM.5 concentration in the air and ensure people's health. Secondly, from the perspective of linear regression model, the author believes that significance test should be carried out on regression coefficients to enhance rigor; in addition, collinearity test should be carried out in multiple linear regression model to eliminate the influence of multicollinearity phenomenon on result accuracy. In the study of predicting PM2.5, the author believes that if short-term prediction is carried out, ARIMA model is better, and if long-term prediction is carried out, models that can maintain long-term dependence relationship of data (such as LSTM model, etc.) are needed. The author hopes that this finding can provide some feasible suggestions for PM2.5 prediction work. The author gives the best application scope of different models. The government and relevant researchers can flexibly adjust the prediction method to improve the accuracy of PM2.5 prediction, facilitate government decision-making and improve people's quality of life.

References

[1] Wang Xuzhi, Zeng Pei, Liu Yonghui. Analysis and Prediction of PM2.5 Concentration in Shanghai. Mathematics in Practice and Understanding, 2017, 47(15):210-217.

[2] Wang Weiduo, Wu Taotao, Zhang Zizhen. PM2.5 Prediction in Hangzhou Based on ARIMA Model. Journal of Harbin Normal University (Natural Science Edition), 2018, 34(03):49-55.

[3] Huang Yun, Jiang Guo, Xu Zhiqian. Application of ARIMA Model in PM2.5 Concentration Prediction in Huangshi City. Journal of Hubei Normal University (Natural Science Edition), 2017, 37(02):38-42.

[4] Zhang Ying, Zhang Jie, Wang Shigong, Kang Ping, Zhang Jiaxi, Zhang Xiaoling, Li Yunchao. Interactive Effects of the Influencing Factors on the Changes of PM2.5 Concentrations in Chengdu,China, Environmental Science,2021,41(10):4518-4528.

[5] Li Jiaxu. Analysis and Prediction of the Influencing Factors of PM2.5 Concentration in Chengdu. Lanzhou University, 2020.

[6] Si Shoukui, Sun Jianliang, Mathematical Modeling Algorithms and Applications, 2015.4

[7] People.cn, URL: http://env.people.com.cn/n1/2021/0805/c1010-32182372.html, last accessed 2021/8/5.

[8] D. Deva HemaJ. TharunG. Arun DevN. Sateesh. A Robust False Spam Review Detection Using Deep Long Short-Term Memory (LSTM) Based Recurrent Neural Network. Journal of computational and theoretical nanoscience, 2020, 17(8).

[9] Zhao Han. Research and Application of Multi-step Prediction Method for Time Series Data Based on RNN. Beijing University of Technology, 2020.

[10] Lu Peng, Liu S, Liu R,et al. Effective long short-term memory with differential evolution algorithm for electricity price prediction. Energy, 2018.