Air quality prediction and early warning model based on LSTM-SARIMA

Depu Zhao †, Senyi Jiang † and Yiman Wu

Wuhan University of Technology, Wuhan, China

† These authors also contributed equally to this work

Abstract. With the development of modern society, air pollution has arisen and people's healthy life is threatened, drawing the attention of contemporary society. In this paper, the influencing factors related to PM2.5 concentration and their importance are obtained analytically by linear interpolation method with the help of multiple linear regression and random forest model with respect to air pollution. Meanwhile, based on LSTM model and SARIMA model, the PM2.5 concentration is weighted to predict the air quality for prewarning. Finally, a multi-step prediction model of AQI is developed, and the air quality index (AQI) is used to measure the air quality condition and explore the factors affecting the AQI value.

Keywords: Random Forest, linear insertion, LSTM, SARIMA, weighted prediction.

1. Introduction

Air quality can reflect the degree of air pollution, which is harmful to human health, ecological environment, and social economy. Its pollution level is affected by many factors, such as PM2.5, PM10, CO, temperature, wind speed, precipitation, etc. The air quality is decreasing day by day during the development of society.

In order to create a healthier living environment, this paper mainly focuses on the factors related to the concentration of pollutants, such as PM2.5 and more accurate prediction of PM2.5 concentration and AQI index, etc. It is important to analyze the factors influencing PM2.5 pollution and the impact of time change, predict the daily air quality warning and effectively develop control strategies.

2. Modeling and empirical analysis

2.1. Detection of factors associated with changes in PM2.5 concentration test

2.1.1. Multiple linear regression analysis modeling

In this study, it is necessary to find the factors related to the variation of PM2.5 concentration, to be able to detect the influence of multiple variables on the dependent variable, and to predict the taken values more accurately. Therefore, the linear interpolation method is adopted to further process the missing values of the data. The linear interpolation method is an interpolation method for one-dimensional data. It is based on the numerical estimation of the left and right proximity of the points to be interpolated in a one-dimensional data series, and the weight is assigned according to the distance to these two points. Suppose we have two known points, which is (x_0, y_0) and (x_1, y_1), and $x_0 < x < x_1$, the linear interpolation method takes the estimated x values by the following formula,

$$y = y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) \tag{1}$$

Multiple linear regression analysis was then used to screen the influencing factors [1]. When set y as the dependent variable, $x_1, x_2, ..., x_n$ as the independent variable, and when the relationship

between the independent variable and the dependent variable is linear, the multiple linear regression model is,

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n * x_n + \varepsilon \tag{2}$$

Including, y is the dependent variable, $x_1, x_2, ..., x_n$ is the independent variable, b_0 is the intercept, $b_1 ... b_n$ is the regression coefficient, and ε is the random error term.

2.1.2. Random forest modeling

It is assumed that the greater the contribution made by an indicator in the growth of a random forest regression tree [2], the greater weight should be assigned to it, and the weight is noted as α. It is calculated as follows:

Step 1: Calculate the mean square error of each node on each tree, where the mean square error of the node s is,

$$MSE_i = \frac{1}{N} \sum_{i=1}^{N_s} (c_s - y_i)^2 \tag{3}$$

Including, N denotes the number of samples, y_i denotes the true value of the samples, and c_s is the model output value.

Step 2: Calculate the nodal contribution of each indicator by expressing the contribution of the indicator j at the node s as the difference in mean square error between the parent and child nodes after branching,

$$V_{js}^{MSE} = MSE_s - MSE_l - MSE_r \tag{4}$$

In the formula, MSE_s and MSE_r are the mean square error of the two sub-nodes after branching, respectively.

Step 3: Calculate the cumulative contribution value of each indicator and define the cumulative contribution of the indicator j as,

$$V_{js}^{MSE} = \sum_{s \in M} V_{js}^{MSE} \tag{5}$$

In the formula, M is the set of nodes t where the indicator j appears in the first regression tree.

Step 4: The weights of each indicator are obtained as,

$$\alpha_j = \frac{\sum_{t=1}^{T} V_{tj}^{MSE}}{\sum_{s \in M} \sum_{t=1}^{T} V_{js}^{MSE}} \tag{6}$$

In the formula, T is the number of regression trees in the forest; m is the set of splitting indicators.

On the other hand, from the results of the model, when the random forest model is trained, randomly resetting the value of each indicator in the dataset in turn, and then observing the change of the model performance. If the model effect decreases much, the indicator has a significant effect on the target value, and it is justified to give a greater weight, and the weight is noted as β [3]. It is calculated as follows:

Step 1: For the trained random forest model selection as a measure of model performance, the calculation formula is,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum i = 1^n (y_i - \bar{y})^2} \tag{7}$$

In the formula, y_i denotes the actual observation, $\hat{y}_i$ denotes the predicted value for that observation, $\bar{y}$ denotes the average of all observations, and n denotes the sample size.

Step 2: The data values of the indicator j are reset and arranged, and the importance scores of j' obtained by calculating the reset and making differences are,

$$R_j = R_l^2 - R_e^2 \tag{8}$$

In the formula, R_l^2 is the performance score before the reset; R_e^2 is the performance score after the reset.

Step 3: Combine all indicators and get the importance score of each indicator. Then the weights of the indicators j are,

$$\beta = \frac{R_j}{\sum_{i=1}^{m'} R_j} \tag{9}$$

In the formula, m' is the number of indicators; R_j is the importance score of indicators.

Finally, the sum is weighted α and β, and combined to obtain the indicator weights as,

$$\varepsilon = \overline{\omega}\alpha_j + (1 - \overline{\omega})\beta_j \tag{10}$$

In the formula, $\overline{\omega}$ is the adjustment factor, takes the value is $0 \sim 1$.

2.1.3. Empirical analysis

(1) Random Forest

Matlab was used to derive and compare the weight of different factors, with PM2.5 as the dependent variable and other influencing factors as independent variables, and the following results were obtained based on random forest (Fig. 1):

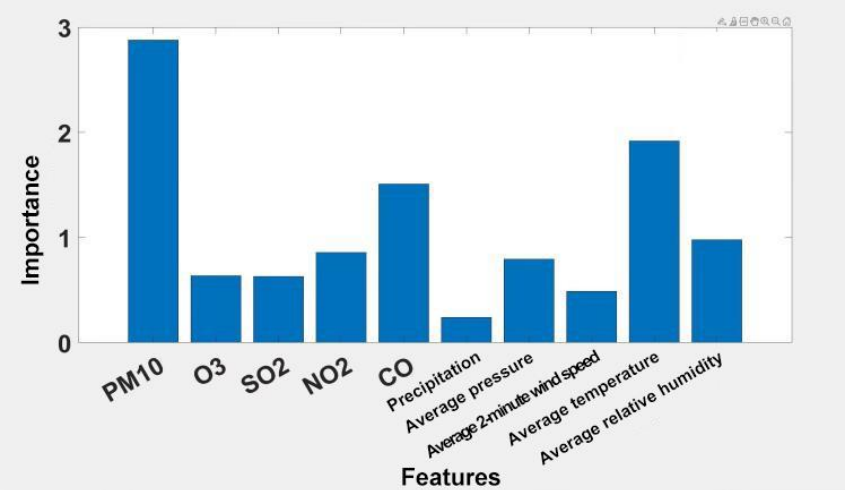

Figure 1. Percentage of the effect of each factor on PM2.5 concentration

It can be obtained that among all the factors, PM10, O3, SO2, NO2, CO, precipitation, average air pressure, average 2-minute wind speed, average air temperature, and average relative humidity have an effect on PM2.5 concentration variation, and among these factors, PM10, average air temperature, and CO have a greater effect on PM2.5, and precipitation has the least effect. The results were evaluated by R2 and obtained R2=0.8684.

At the same time, the top three influencing factors with the largest share were taken and modeled again, that is the results as in Fig. 2 were obtained:

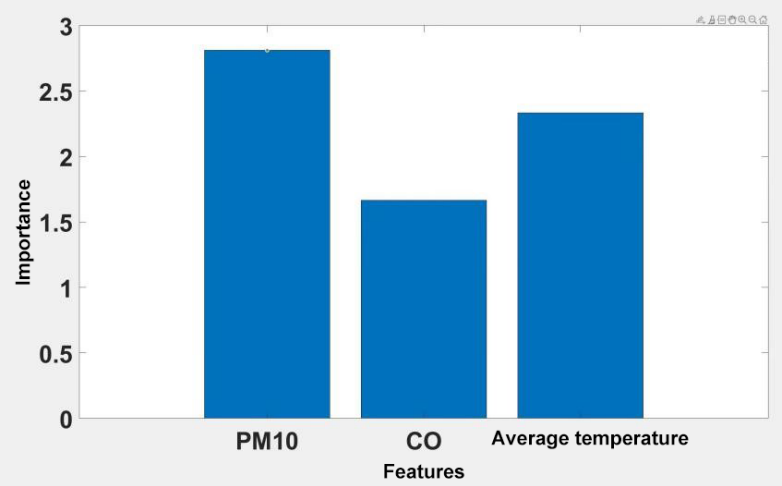

Figure 2. Percentage effect of PM10, NO2, and average air temperature on PM2.5 concentration

The R2=0.84464 is obtained, and when the influence of other factors on the change of PM2.5 concentration is removed and only the influence of the most significant factor is considered, the R2 does not change significantly and is basically stable, so that it can be accurately concluded that PM10, average temperature and CO have a greater influence on PM2.5, while precipitation, average two-minute wind speed and average air pressure have a smaller influence on its concentration change. is smaller.

Multiple linear regression analysis

(2) Multiple linear regression analysis

In the multiple linear regression analysis, the relevant data were analyzed and processed using SPSS, and the data on the weight of different influencing factors were obtained as shown in Table 1:

Table 1. Percentage effect of PM10, NO2, and average air temperature on PM2.5 concentration

PM10	CO	Average temperature	NO2	Precipitation
0.622	0.295	-0.255	-0.121	-0.033
Average relative humidity	Average two-minute wind speed	SO2	Average air pressure	O3
0.028	-0.025	-0.007	0.006	-0.004

The analysis of the results of the test F can be obtained, that the significance value P is 0.000, which presents significance at the level and rejects the original hypothesis that the regression coefficient is 0. Therefore, the model basically meets the requirements. For the performance of variable co-linearity, all of VIF are less than 10, so the model has no problem of multiple co-linearity and the model is well constructed.

The equation of the model is as follows: y =-16.804+0.444 * PM10-0.003 * O3-0.04 * SO2+31.349 * CO-0.179 * NO2-0.215 * Precipitation+0.017 * average air pressure-0.797 * average 2-minute wind speed+ 0.076 * average relative humidity-0.787 * average temperature.

Removing the influences with standardized coefficients below 0.1 leads to the following Table 2:

Table 2. Table of main influencing factors

Influencing Factors	PM10	Average temperature	CO
Standardization factor	0.622	0.256	0.295

In summary, for both models, the main influencing factors point roughly to PM 10, average temperature, CO.

2.2. Multi-step prediction test of PM2.5 concentration

2.2.1. SARIMA's time series forecasting model equation

SARIMA, fully known as the seasonal difference autoregressive moving average model, is an extension of the ARIMA model. ARIMA is a classical time series forecasting method that represents an autoregressive moving average model [4]. Unlike the ARIMA model, the SARIMA model is specifically designed to deal with time series with significant seasonal factors. The model developed by transforming a non-stationary time series into a stationary time series and then regressing the dependent variable only on its lagged values and on the present and lagged values of the random error term, the formula defines as,

$$Y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{11}$$

The expression of the SARIMA model is generally,

$$SARIMA(p,d,q)(P,D,Q)m \tag{12}$$

Including, p is the autoregressive order, q is the moving average order, d is the number of differences of the time series, P, Q and D denote the order of seasonal autoregressive, seasonal moving average and seasonal difference, respectively, then m denotes the seasonal period.

2.2.2. LSTM-SARIMA weighted prediction model

Weighted forecasting with different forecasting methods refers to combining multiple forecasting methods for forecasting. This method uses a weighted average of the forecasts from multiple forecasting models to produce a more accurate and robust forecast. Optimization can usually be performed by methods such as cross-validation to obtain the best results. Let $\hat{y}_i (i = 1,2,...,m)$ be the prediction result y'_{ij} of the same problem obtained by m prediction methods (each method passing its own test), wihch is the simulated value of the $i(i = 1,2,...,m)$ prediction method on the original data y_i,

$$J = \sum_{i=1}^{m} x_i \hat{y}_i \tag{13}$$

Including, $\sum_{i=1}^{m} x_i = 1, x_i \geq 0 (i = 1,2,...,m).$

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \tag{14}$$

Then, with reference to these coefficients, x_i as an optimizing idea, which can find x_i satisfying that,

$$\min L = \sum_{j=1}^{n} x^T E_j x \tag{15}$$

We use LSTM and SARIMA two models for forecasting, respectively, so we weight the forecasts obtained from both to obtain the LSTM-SARIMA weighted forecasting model with the following equation:

$$\min L = \sum_{j=1}^{12} x^T E_j x \tag{16}$$

Including, $\sum_{i=1}^{2} x_i = 1, x_i > 0 (i = 1,2)$, x_1 is the weight parameter of the prediction result obtained using the LSTM model and x_2 is the weight parameter of the result obtained using the SARIMA model. And the expression of the weighted predicted values is as follows:

$$J_t = x_1 y_{1t} + x_2 y_{2t} \tag{17}$$

Including, y_{1t}, y_{2t} denote the predicted values of LSTM and SARIMA at moment t, respectively, and J_t denotes the weighted average of the two.

2.2.3. Root mean square error (RMSE) effect evaluation

The root means square error, also known as the standard error, is the arithmetic square root of the mean square error. The mean square error has a different scale from the data scale, which does not visually reflect the degree of dispersion, so the square root is opened on the mean square error to obtain the root mean square error,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{18}$$

Including, y_i denotes the true observed value of the i-th sample, $\hat{y}_i$ denotes the value of the i-th sample predicted by the model, and n denotes the number of samples. A smaller $RMSE$ means a better model prediction, and a larger $RMSE$ means a larger error.

In this study, the assessment of the $RMSE$ effect of asynchrony can be obtained as shown in Table 3:

Table 3. Predicted $RMSE$ results of PM2.5 concentration

Predicted step-size	Three-step prediction	Five-step forecast	Seven-step forecast	Twelve-step forecast
RMSE	19.5612	23.0424	23.1805	24.3436

As can be seen from Table 2, the $RMSE$ value increases with the increase of the prediction step, and the prediction effect of the model slightly decreases; therefore, the three-step prediction should be preferred.

The predicted PM2.5 concentration results are shown in Table 4:

Table 4. Predicted PM2.5 concentration results

Date (month/day/year)	Apr. 30th, 2023	May 1st, 2023	May 2nd, 2023	May 3rd, 2023	May 4th, 2023	May 5th, 2023
PM2.5	34	41	48	58	67	66
Date (month/day/year)	May 6th, 2023	May 7th, 2023	May 8th, 2023	May 9th, 2023	May 10th, 2023	May 11th, 2023
PM2.5	55	53	50	44	52	64

2.3. Root mean square error *RMSE* test

2.3.1. *RMSE* effect evaluation and visualization prediction

After building the LSTM-SARIMA weighted multi-step forecasting model, it was trained and parameters were selected and weighted forecasts were made (see Fig. 5), and the final optimization yielded an *RMSE* =24.7872, which performed relatively well. It was then applied to the test set to predict the AQI values for the next 12 days, as shown in Table 5 and Fig. 3 below,

Table 5. Predicted PM2.5 concentration results

Date (month/day/year)	Apr. 30th, 2023	May 1st, 2023	May 2nd, 2023	May 3rd, 2023	May 4th, 2023	May 5th, 2023
AQI	123	103	104	74	63	69
Level-color of prewarning	Blue	Blue	Blue	None	None	None
AQI	78	79	88	76	80	81
Level-color of prewarning	None	None	None	None	None	None

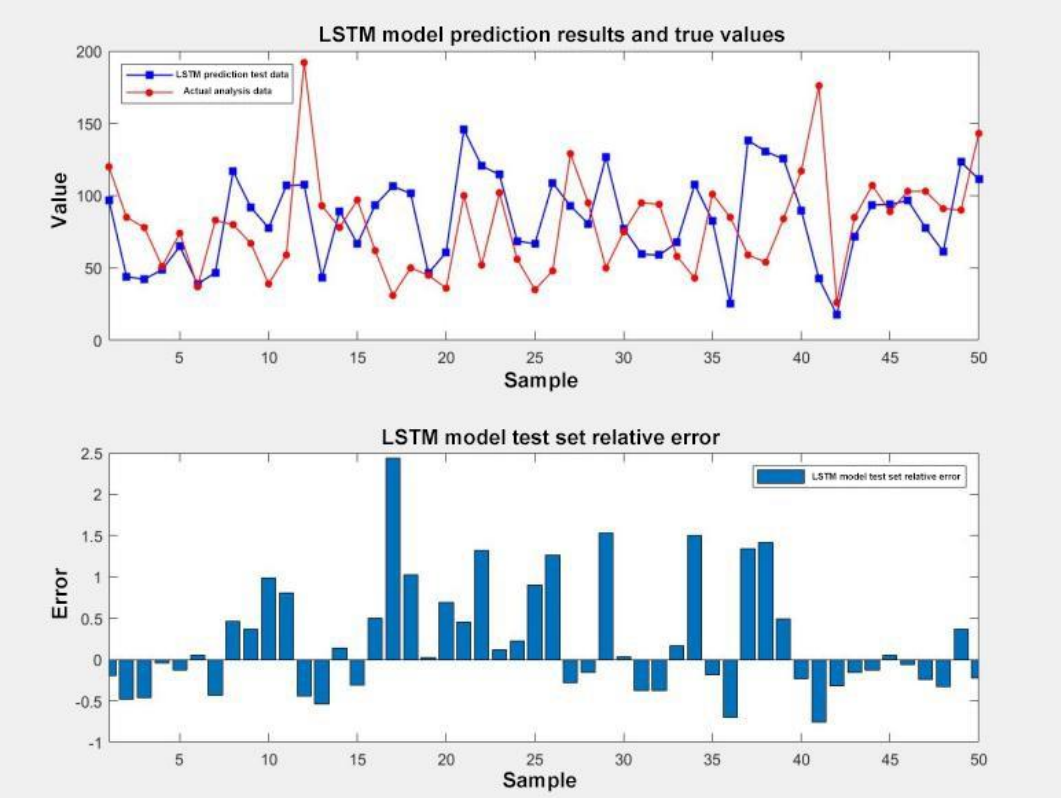

Figure 3. Schematic diagram of AQI prediction results

It can be seen that the fitting trend is roughly consistent and the relative error is small; the alarm level colors are shown in Table 6, in which the air quality is slightly polluted in the first 3 days and relatively good in the last 9 days.

Table 6. Summary of the number of level-color of prewarning

Level-color of prewarning	Blue	Yellow	Orange	Red	None
Number of days (days)	3	0	0	0	9

3. Conclusion

In order to improve and perfect the response and disposal mechanism of heavy polluted weather, it is necessary to improve people's awareness of heavy polluted weather, response ability and society's environmental refinement management level. When issuing emergency plans for polluted weather, it is necessary to strengthen monitoring and early warning and energy conservation and emission reduction to minimize the impact of polluted weather.

References

[1] Zhao Zheng, Yu Xiaojie, Xiong Yuzheng et al. PM2.5 prediction model based on regression analysis and decision tree algorithm [J]. Changjiang Information &Communications, 2022, 35 (11): 9 - 11.

[2] Xia Weihuai, Liu Jiali, Feng Fenling. Demand forecast of railway refrigerated transportation based on random forest [J]. Journal of Railway Science and Engineering, 2022, 19 (4): 909 - 916.

[3] Wu Weijie, Zhang Jingxiang. Random Forest Feature Selection Algorithm Based on Categorization Information and Application [J]. Computer Engineering and Applications, 2021, 57 (17): 147 - 156.

[4] Zhang Min. Research on the Development Trend of College Students' Physical Quality Based on Time Series Analysis [D]. North University of China, 2020.