ISSN: 2710-0723

Prediction and Early Warning of Air Quality based on the LSTM-ARIMA Model

Zishan Li, Yan Zhang, and Yida Wang*

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233030, China

Abstract

At present, China's economy is transforming from a stage of high-speed growth to a stage of high-quality development. Building an ecological civilization is an important part of the realization of the Chinese dream of national rejuvenation. Air pollution will cause harm to human health, ecological environment, social and economic aspects, and its pollution level is affected by many factors, such as PM2.5, PM10, CO, temperature, wind speed, precipitation and so on. In order to implement the party's 20th spirit, strengthen the coordinated control of pollutants, basically eliminate heavy pollution weather, improve and improve the response and disposal mechanism of heavy pollution weather, a place issued an emergency plan for pollution weather, which will strengthen monitoring and early warning, energy conservation and emission reduction, and minimize the impact of pollution weather. To explore the influencing factors of PM2.5 concentration change, more accurate prediction PM2.5 concentration and AQI index, this paper for the prediction and warning of air quality, using the daily pollutants in January 2015 to April 2023, the data of analysis of PM2.5 concentration, the importance of the top three factors PM10, average temperature, CO as auxiliary variables, to construct the LSTM-ARIMA combination model. First, LSTM was used for multi-factor prediction. In order to improve the accuracy of the prediction results, the prediction error of LSTM was then linearly corrected based on the ARIMA model, the time step was adjusted to observe the previous data, and the prediction results of the model were evaluated by root mean square error (RMSE).

Keywords

PM2.5 Calculate; AQI Calculate; The LSTM-ARIMA Combined Model; MATLAB.

1. Introduction

1.1. Background Knowledge

In order to deal with heavy pollution weather and improve the prevention and emergency response capacity, it is also necessary to implement fine environmental management. For heavy pollution weather, the release of pollution weather emergency plan is an important measure, aiming to minimize the impact of pollution weather on human beings and the environment. In order to effectively prevent and control air pollution, it is necessary to quantify the air quality and find its change law. The Air Quality Index (AQI) is a measure of the quality of life. The government, all sectors of society and individuals should make joint efforts to actively respond to relevant policies and take various measures to protect the environment and reduce the impact of air pollution. However, despite the existing air quality monitoring system, it still needs to be improved. There is a lag in real-time monitoring, which cannot predict the air pollution situation in advance and provide reliable early warning information to the government in time.

ISSN: 2710-0723

At present, the methods of air quality prediction are mainly as follows: first, the air quality potential prediction method considering weather conditions and meteorological parameters, where the reference condition is single and the accuracy is not high; the second, the linear and nonlinear laws based on the historical data of specific regions and weather conditions.

1.2. Research Meaning

With the acceleration of socialization process, air pollution phenomenon is becoming more and more serious, for human health, ecological environment, social development caused certain harm, the pollution level affected by pollutants and meteorological changes, and many other factors, such as PM2.5, PM10, CO, temperature, wind speed, precipitation, etc., to explore these factors, and the AQI index more accurate prediction, is a common concern of the scientific community and policy makers.

Air quality testing is of great significance for the management and improvement of indoor air quality. The so-called air quality detection is to conduct the timing and continuous sampling and measurement of the pollutants existing in the air. Through the air quality detection, the change law and development trend of air quality can be carried out accordingly, and the smooth development of air pollution forecast work can be promoted, providing theoretical basis for government departments to implement relevant environmental protection laws and regulations and carry out environmental quality control.

1.3. Literature Review

For the construction of weather prediction model, the prediction model based on statistics is the historical number of specific regions and weather conditions.According to the accumulation on the basis of the meteorological forecast, Liu (2015) combined ARIMA, BP and exponential smoothing models, determined the weight of each model according to the entropy weight method, and applied the combined model to the actual prediction. Compared with a single model, the results found that the accuracy of the combined model was better than that of the single model [1]. Yang and Jian (2017) used signal decomposition, namely differential evolution, to improve the Elman neural network model, and improved the accuracy of the model[2]. On the basis of the traditional SVM model, Ni Zhiwei et al. (2016) adopted the artificial fish method to optimize the original model, which has high stability and credibility [3]. Peng Yi (2020) proposed the ARIMA-SVM combination prediction method, and finally verified that this method works better[4]. The study of statistical prediction methods by several experts and scholars increases the possibility of statistical prediction methods, improves the accuracy of air quality prediction, and provides a solid foundation for the subsequent air quality.

2. Model Introduction

2.1. LSTM Model

LSTM neural network is a variant of recursive neural network (RNN). LSTM structure [5] adopts the control gate mechanism, which is composed of memory cell, input gate, output gate and forgetting gate. Its core concept lies in the state of the cell and the structure of the gate. The cell state can transmit the relevant information in the sequence process, which overcomes the influence of short-term memory and uses the structure of the gate to determine which information should be saved or forgotten.

The input gate i_t is used to determine how much input x_t of the current moment can be saved to the cell state C_t. The input gate it receives the input x_t of the current moment and the output h_t of the previous moment as the input, first calculates by using the *sigmoid* activation function, and later using the *tanh* activation function to obtain new candidate memory cells to calculate the $\tilde{c}_t$. The calculation formula is as follows:

ISSN: 2710-0723

$$x_t = \sigma(W_{ix}x_t + W_{ih}h_{t1} + b_i). \quad (1)$$

$$\tilde{c}_t = tanh(W_{\tilde{c}_t x}x_t + W_{\tilde{c}_t h}h_{t1} + b_{\tilde{c}_t}). \quad (2)$$

The forgetting gate f_t is used to assess the importance of memory information at the current moment, the forgetting gate of 0 means that any information of C_{t1} is not transmitted to C_t, and a value of 1 means that all information is transmitted to C_t.

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f). \quad (3)$$

After the processing of the forgetting gate and the input gate, the cell state is updated to form a long-term memory. The update formula is:

$$C_t = f_t C_{t-1} + i_t \tilde{c}_t. \quad (4)$$

The output gate o_t is used to selectively output the cell state at the current moment, and the final output state is h_t:

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o). \quad (5)$$

$$h_t = o_t tanh(C_t). \quad (6)$$

Where, W_{ix}, W_{ih}, $W_{\tilde{c}_t x}$, $W_{\tilde{c}_t h}$, W_{fx}, W_{fh}, W_{ox}, and W_{oh} represent the weight matrix of each control gate, b_i, $b_{\tilde{c}}$, b_f, and b_o represent the bias of each control gate, and are the *sigmoid* activation function and *tanh* activation function, respectively.

2.2. ARIMA Model

ARIMA (p, d, q) model [6] is a time series prediction method that can predict future values based on historical and current values. It is actually a combination of AR model and MA model. Firstly, check the time series stability, determine the non-stationary sequence parameter d by stabilizing the difference, and ensure that the data sequence is a stable non-white noise sequence; secondly, the autoregressive parameter p and the moving average parameter q are determined by drawing the partial autocorrelation map (PACF) and autocorrelation map (ACF) to establish ARIMA (p, d, q).

2.3. LSTM-ARIMA Combined Model

Firstly, predict the long-short memory neural network (LSTM) model as auxiliary variables, and the prediction error of the LSTM model is made linearly corrected, based on the autoregressive difference moving average (ARIMA) model to ensure more accurate prediction results; evaluate the prediction error with the root mean square error (RMSE) index, and finally, the two are combined to obtain the final value of the LSTM-ARIMA combination model.

The idea of the LSTM-ARIMA combination model is shown below:

ISSN: 2710-0723

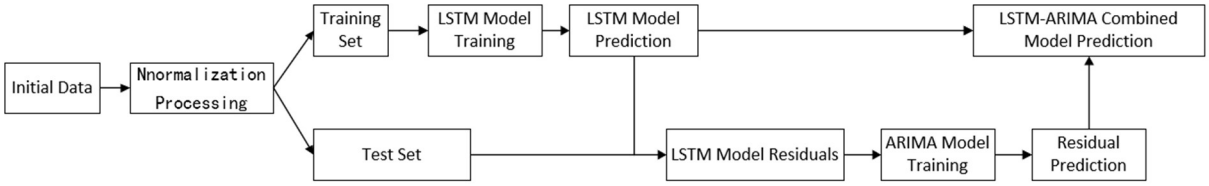

Figure 1. Train of thought

Taking the predicted PM2.5 concentration as an example, the specific modeling steps are described as follows:

1) The data were analyzed and processed, and the correlation analysis was conducted to select the factors related to the change in PM2.5 concentration, and the importance of the factors on PM2.5 concentration was explored using the random forest regression model.

2) The top three data on the degree of effect after normalization were used as auxiliary variables and the PM2.5 concentration itself as input, and then the hyperparameters in the LSTM model were optimized by random search method to establish the optimal prediction model.

3) Take time t as an example, make the true value of PM2.5 and concentration yt and the predicted value of LSTM model y ˆ t, and find the prediction error value at that time, that is, formula (7):

$$\hat{e}_t = y_t - \hat{y}_t. \tag{7}$$

4) ARIMA is established according to the residual sequence obtained by the LSTM model, the model uses ACF to test the stationarity of the sequence, observe the autocorrelation map and partial autocorrelation map to estimate the model parameters, the AIC value under different parameters is calculated, the optimal prediction model is found, and the output prediction error value $\hat{e}_t$ is obtained.

5) The linear combination of the LSTM neural network model and the ARIMA model output results to obtain LSTM-ARIMA final prediction result $\hat{Y}_t$ of the LSTM combined model engraved at this time:

$$\hat{Y}_t = \hat{y}_t + \hat{e}_t. \tag{8}$$

3. Empirical Analysis

3.1. Screening Variables

3.1.1. Correlation Analysis

The factors related to the change of PM2.5 concentration, including five pollutants PM10, O_3, SO_2, NO_2 and CO, as well as five meteorological factors: precipitation, average air pressure, average wind speed in 2 minutes, average temperature and average relative humidity. Based on the Pearson's correlation coefficient to explore the correlation between the data features.

ISSN: 2710-0723

Figure 2. Thermal map of the correlation coefficient

3.1.2. Build a Random Forest Regression Model

Further fitting the random forest regression models to calculate the importance of each independent variable to the dependent variable PM2.5. Samples are randomly collected from the original data set, and a training set is generated for each decision tree. The random forest consists of multiple decision trees. The prediction result of the regression model is the mean of the predicted value of all decision trees, and the percentage is the influence degree of factors affecting PM2.5 concentration according to the following table:

Table 1. The extent to which each factor affected the PM2.5 concentration

Sort	Factor	Importance score
1	PM10	0.713120
2	Average temperature of air	0.100150
3	CO	0.094852
4	Average wind speed in 2 minutes	0.020583
5	Average relative humidity	0.017494
6	NO_2	0.015668
7	SO_2	0.013012
8	Average air pressure	0.011325
9	O_3	0.713120
10	Precipitation	0.100150

3.2. Normalization

In order to eliminate the influence of different attribute data dimensions, this paper chooses to use the Min-Max normalization method to transform the original data and map the data between [0,1] to eliminate the influence of the dimension. The change formula of the Min-Max normalization method is:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{9}$$

ISSN: 2710-0723

The top three influence factors on the degree of PM2.5 concentration PM10 concentration, mean air temperature, CO concentration and PM2.5 concentration were selected for Min-Max normalization.

3.3. LSTM Model Building

3.3.1. Hyperparameter Optimization

This paper uses normalized PM10 concentration, mean air temperature, CO concentration data as influencing factors as well as PM2.5 concentration itself as input, and the top 80% of samples of the data as the training set. The prediction step size of the model was set to 3, and the MSE loss function, the tanh activation function, and the Adam algorithm optimizer were selected. The Batch _ size of the experiment was 64,100 training iterations, the number of neurons was 150, the discard rate was 0.5, and the learning rate was 0.01. In order to improve the learning effect of the model, the hyperparameters in the LSTM model were optimized by random search method (Random-izedSearch CV) to find the minimum verification loss.

Table 2. Optimal hyperparameters

Parameter name	Scope of reference	Optimal parameters
the-units	[50, 100, 150]	100
dropout rate	[0.2, 0.5]	0.5
Ir-rate	[0.001, 0.01]	0.01
epochs	[30, 50, 100]	100
batch-size	[32, 64, 128]	128

According to Figure 3, we can see that the loss drops rapidly and converges on the training set, and the loss of the model on the test set is also constantly decreasing and finally stabilizing. After convergence, the verification loss and the training loss basically coincide, and the model has no overfitting or underfitting phenomenon, and the fitting effect is better.

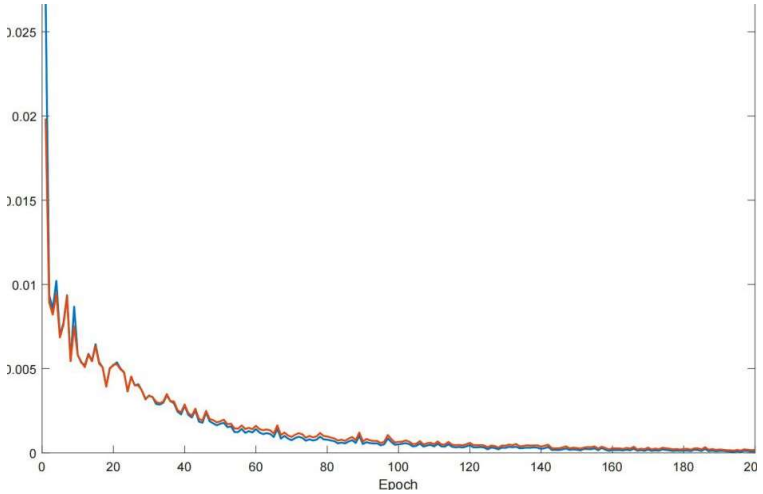

Figure 3. The LSTM model loss change plot

3.3.2. LSTM Model Test Set Fit

The PM2.5 concentration was set to 3 to predict the prediction, and the prediction effect was shown in Figure 4. Similarly, the normalized AQI air quality index itself was used as input to predict the effect in Figure 5.

ISSN: 2710-0723

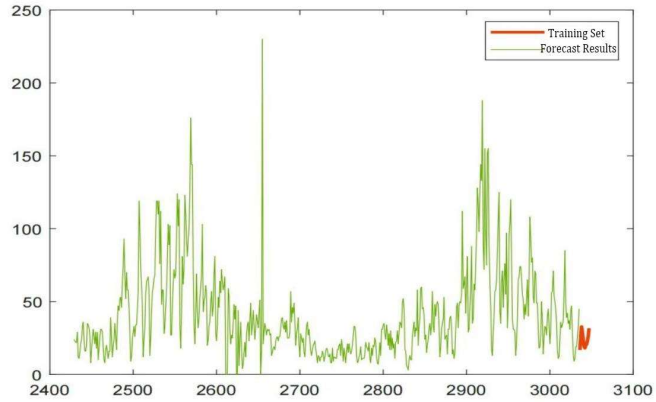

Figure 4. The LSTM model for PM2.5 concentration prediction

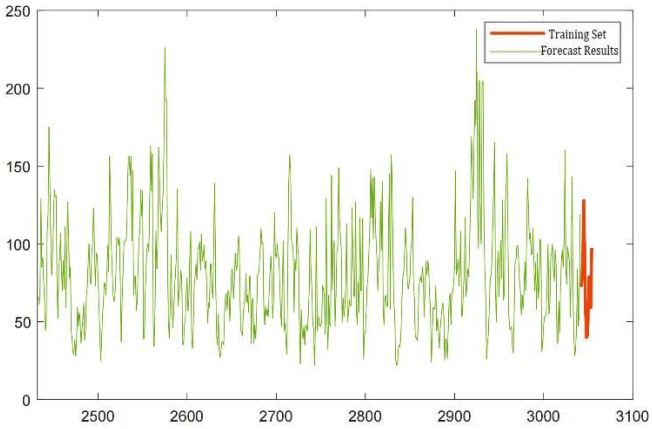

Figure 5. The LSTM model for AQI prediction

3.4. ARIMA Model Building

3.4.1. Stabilization and White Noise Test

First deal with the residual sequence of stability and white noise test, to PM2.5 concentration data, for example, according to the residual timing diagram and autocorrelation diagram, the sequence has no obvious linear growth trend, and in addition to the delay order 3 autocorrelation coefficient in two times the standard deviation, other autocorrelation coefficient are within two times the standard deviation, the speed of autocorrelation coefficient decay to zero is fast, can be considered the sequence has short-term correlation, therefore, can judge the sequence is smooth sequence.

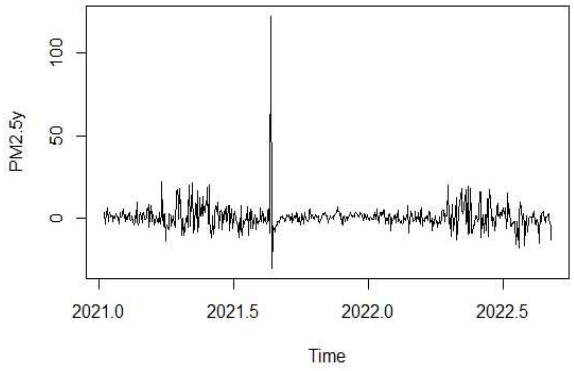

Figure 6. Time sequence diagram of residuals

ISSN: 2710-0723

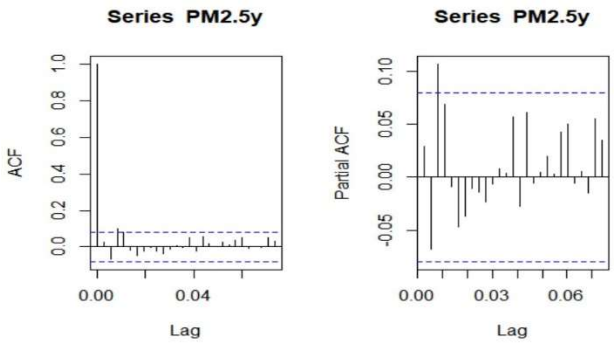

Figure 7. ACF and PACF Figure

Later, the sequence is tested by white noise test, that is, the null hypothesis that there is no correlation between the sequence values of lagging m is established, and the p-value at lag 6 is 0.021, less than 0.05, and the null hypothesis can be rejected. Therefore, the sequence can be determined as a stationary non-white noise sequence.

3.4.2. ARIMA Model Order and Prediction

Observe Figure 7, the autocorrelation graph is 3 order censoring, and the partial autocorrelation graph is 2 order censoring, so the model parameters can be estimated as ARIMA (3,0,2), ARIMA, ARIMA (2,0,3), ARIMA (3,0,1) models, and the AIC values are 4153.74,4153.85 and 4154.3. According to the minimum information principle, the ARIMA (3,0,2) model is established. Similarly, the ARIMA (5,0,1) model was established for AQI to predict. The model significance and parametric tests of the two models were conducted respectively, and the visual results of predicting the sequence values of the next 12 periods are shown in the figure:

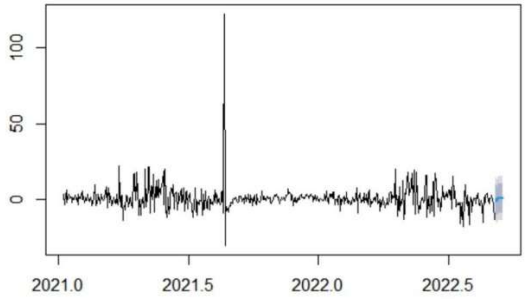

Figure 8. ARIMA Model to predict the PM2.5 concentration results

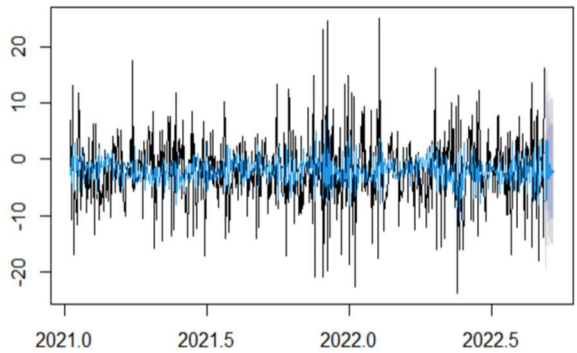

Figure 9. ARIMA Model predicts the AQI index results

ISSN: 2710-0723

3.5. LSTM-ARIMA Model Establishment and Prediction Results

The results $\hat{y}_t$ output by the LSTM neural network model and the ARIMA output $\hat{e}_t$ are linear combined $\hat{Y}_t = \hat{y}_t + \hat{e}_t$, to obtain the final prediction result $\hat{Y}_t$ engraved by the LSTM-ARIMA combined model at this time:

Table 3. The PM2.5 concentration prediction results

Date	$\hat{y}_t$	$\hat{e}_t$	$\hat{Y}_t$
2023/4/30	19.419804	0.3242682	19.7440722
2023/5/1	30.596958	0.9457892	31.5427472
2023/5/2	36.50737	-0.8167717	35.6905983
…	…	…	…
2023/5/9	24.245199	0.9618557	25.2070547
2023/5/10	27.814337	0.9750108	28.7893478
2023/5/11	34.813389	0.9347997	35.7481887

The prediction effect of 3,5,7, and 12 steps was evaluated by root mean square error (RMSE). The smaller the RMSE, the better the prediction effect of the model will be. By comparing the RMSE predicted by the ARIMA-LSTM model, we show that the RMSE predicted in 3 steps is the smallest and the best prediction effect is optimal. Therefore, this combined model is used to predict the next 12 periods PM2.5 concentration, the results are shown in the table below:

Table 4. The PM2.5 concentration prediction results

Prediction step	3 steps	5 steps	7 steps	12 steps
RMSE	6.418161	7.819365	7.463501	7.007037

Data	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
PM2. 5	19.7441	31.5427	35.6906	30.9546	24.4801	19.9620
Data	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
PM2. 5	18.6631	21.2931	22.0164	25.2071	28.7893	35.7482

Figure 10. The PM2.5 prediction results

Similarly, the above model is used to predict the AQI of the next 12 periods, and the early warning level of daily air quality is given according to the division of early warning levels. The forecast results are shown in the figure below:

Data	2023/4/30	2023/5/1	2023/5/2	2023/5/3	2023/5/4	2023/5/5
AQI	44. 5622	82. 8214	135. 3216	95. 8716	54. 1911	37. 9113
Warning Color Level	Not have	Not have	Not have	Not have	Not have	Not have
Data	2023/5/6	2023/5/7	2023/5/8	2023/5/9	2023/5/10	2023/5/11
AQI	36. 8563	48. 5879	77. 3571	62. 2952	57. 7828	100. 3205
Warning Color Level	Not have	Not have	Not have	Not have	Not have	Blue

Figure 11. The AQI prediction results

ISSN: 2710-0723

Warning Color Level	Blue	Yellow	Orange	Red	Not have	Total
Days	1	0	0	0	11	12

Figure 12. Summary of the early warning level color times

4. Conclusion

(1) Based on the air quality data of PM2.5 concentration and AQI index from 2014 to April 2023, and in order to verify the universality and practicability of the model, the data were divided into training set and prediction set for fitting and prediction.If you follow the “checklist” your paper will conform to the requirements of the publisher and facilitate a problem-free publication process.

(2) The LSTM-ARIMA combined model predicted the air quality quite well. First, the LSTM model was used to predict the prediction results of multiple influencing factors on the dependent variables, and then the prediction error of the LSTM model was corrected linearly based on the ARIMA model to make the prediction results more accurate.

(3) PM2.5 concentration and AQI index may show significant changes in different seasons, there are certain seasonal influence, after sensitivity analysis, the model can adapt to the emergence of different seasonal patterns, for periodic change sequence have better fit and prediction effect, model sensitivity is higher, can better to capture the seasonal pattern, for seasonal data have good fitting and prediction ability.

References

[1] Liu D J, Li L. Application Study of Comprehensive Forecasting ModelBased on Entropy Weighting Method on Trend of PM2.5 Concentration inGuangzhou, China [J]. International journal of environmental research and publichealth, 2015,12(6):7085-7099.

[2] Yang Z S, Wang J. A new air quality monitoring and early warningsystem: Air quality assessment and air pollutant concentration prediction[J]. Environmental Research, 2017, 158:105-117.

[3] Ni Zhiwei, Zhu Xuhui, Cheng Meiying. .Air Quality Prediction Method Based on Fish Swarm and Fractal Dimension[J]. Pattern Recognition and Artificial Intelligence, 2016,29 (12): 1122-1131.

[4] Yang Taofeng,Peng Yi. A hybrid ARIMA-SVM model for the study of air quality prediction based on

[5] improved PSO[J]. Journal of Yunnan University (Natural Science Edition), 2020,42 (05): 854-862.

[6] He K,Zhang X,Ren S,et al.Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition [C]// IEEE Transactions on Pattern Analysis & Machine Intelligence,2014: 1904-16.

[7] Wang Xia,Chen Xiaojian. SPATIAL-TEMPORAL CHARACTERISTICS AND CONSTRAINT OF PRIMARY SCHOOL COMMUTING IN XI’AN CITY[J]. Urban Planning, 2018,42 (11): 148-158.