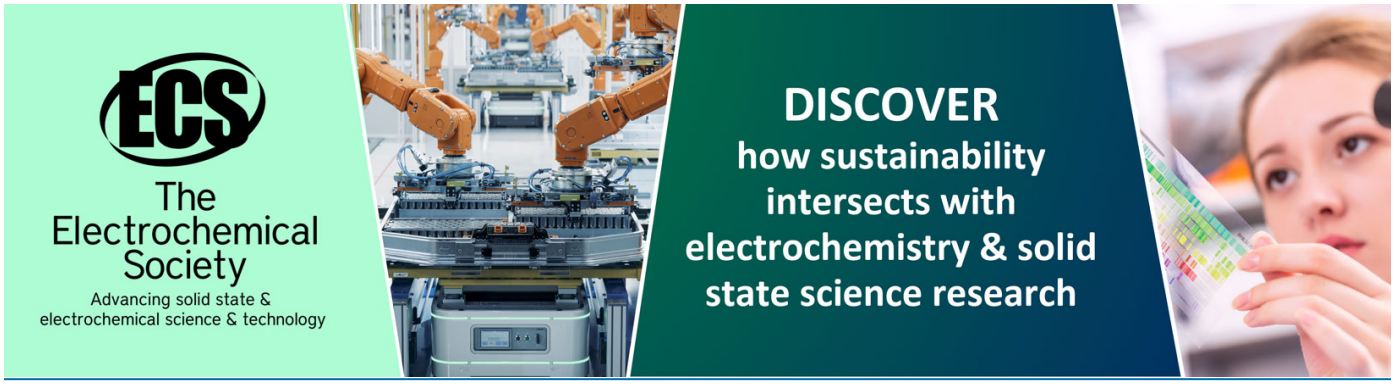
Prediction of PM2.5 concentration in Guangxi region, China based on MLR-ARIMA



Prediction of $PM_{2.5}$ concentration in Guangxi region, China based on MLR-ARIMA

Pengzhi Wei[1], Shaofeng Xie[1,2*], Liangke Huang[1,2], Ge Zhu[1], Youbing Tang[1] and Yabo Zhang[1]

[1] College of Geomatics and Geoinformation, Guilin University of Technology, Guilin, Guangxi, 541006, China

[2] Guangxi Key Laboratory of Spatial Information and Geomatics, Guilin, Guangxi, 541006, China

Author's email: chgcwpz@163.com

*Corresponding author's email: xieshaofeng@glut.edu.cn

Abstract: In recent years, the problem of atmospheric pollution has received more and more attention. Combining the concentration data of various air pollutants monitored by the air quality monitoring stations in Nanning, Guilin, and Baise in Guangxi province in 2017 and the precipitable water vapor (PWV) data obtained by sounding stations in the three cities, analyzed the changes of $PM_{2.5}$ and PWV in major cities in Guangxi and build the multiple linear regression-differential autoregressive moving average (MLR-ARIMA) models respectively make short-term predictions for the changes in $PM_{2.5}$ concentration in the three cities. Among them, the mean absolute error (MAE) of the prediction results of Nanning, Guilin and Baise are 7.57μg/m^3, 12.75μg/m^3 and 7.67μg/m^3, compared with the multivariate linear regression model and the neural network model, the prediction accuracy of this model in Nanning is 43.55% and 46.50% higher than that of the multiple linear regression model and neural network model, respectively, and in Baise is 21.41% and 26.32% higher accordingly, The model prediction effect in Guilin is optimal for the neural network model, which improves 24.46% and 11.84% compared with MLR and MLR-ARIMA models, respectively, where MLR-ARIMA model still has 14.31% accuracy improvement compared with MLR model. This study has some reference value for $PM_{2.5}$ prediction work in major cities in Guangxi, China.

1.Introduction

The problem of atmospheric haze pollution has received more and more attention in recent years, and scholars at home and abroad have conducted a series of studies on the prediction and prevention of $PM_{2.5}$. Yan Zuoning et al [1] used an ARIMA model to make short-term predictions of $PM_{2.5}$ concentrations in Shenzhen, China. Asha B. Chelani et al [2] used a combined multiple linear regression and autoregressive model with meteorological parameters to complete $PM_{2.5}$ concentration predictions for five cities in the Indian region. Doreswamy et al [3] developed a machine learning prediction model using the 2012-2017 air quality inspection dataset in Taiwan as an experimental sample.Zhao Yun et al [4] used the whale optimization algorithm as well as the wolf pack algorithm to mix and optimize BP neural networks to build a WPA-WOA-BP neural network model and predict $PM_{2.5}$ concentrations in Guilin city; Li Jianxin et al [5] used the air quality and meteorological data of Ganzhou city for the whole year of 2017 to build the MRMR-HK-SVM model. The experimental

results showed that the MRMR-HK-SVM model has better generalization ability and can predict $PM_{2.5}$ concentration more accurately compared with the traditional SVM model.

In recent years, with the development of satellite technology, more and more scholars have applied satellite products to the prediction and prevention of $PM_{2.5}$. Zhang et al [6], Zhou et al [7] conducted regional PWV and $PM_{2.5}$ correlation studies based on Beijing and Wuhan respectively, and both of them showed good correlation.Guo et al [8] established a $PM_{2.5}$ concentration prediction method based on random forest algorithm considering GNSS meteorological parameters based on Beijing Fangshan station data, and established a $PM_{2.5}$ random forest prediction model incorporating GNSS meteorological parameters, with good results in a certain accuracy range.Wang et al. have studied the influence of water vapor and wind speed on haze variation in Beijing [9], and the correlation between GPS water vapor and $PM_{2.5}$ mass concentration observation data in Hebei Province [10], and correlated $PM_{2.5}$ concentration with atmospheric pollutants, GNSS PWV and wind speed, and used BP neural network to construct urban $PM_{2.5}$ concentration model and regional $PM_{2.5}$ concentration model by combining these types of factors [11].

The above studies conducted some correlation analysis and established corresponding $PM_{2.5}$ concentration prediction models for different regions combined with different meteorological factors and PWV and $PM_{2.5}$, and all achieved good results. However, there is no better secondary processing for the data residuals generated after the model prediction, and there is a lack of research for further exploration of the model accuracy. Because the causes of $PM_{2.5}$ are complex and have obvious spatial and temporal heterogeneity, there is no one type of $PM_{2.5}$ prediction model that can be applied to all regions.Therefore, this paper focuses on Nanning, Guilin and Baise cities in Guangxi region as the study area, and its 2017 urban air quality monitoring station data set and Precipitable Water Vapor (PWV) data obtained from sounding stations as the basis to explore the variation pattern and correlation between PWV and $PM_{2.5}$ in Guangxi region, establish various prediction models and The ability of short-term prediction in the region is compared and analyzed.

2.Materials and Methods

2.1.Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical analysis method that uses regression analysis to determine quantitative relationships among multiple variables that are interdependent. A regression analysis that generally includes two or more independent variables and shows a linear relationship between the dependent and independent variables is called a multiple linear regression analysis [12].

$$y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i (i = 1,...,n) \quad (1)$$

Where： X_{i1} is the 1st explanatory variable for individual i; X_{i2} is the 2nd explanatory variable for individual i; k is the number of explanatory variables; ε_i is a perturbation term; β is the regression coefficient.

2.2.Neural Network Model

The neural network model includes an input layer, a hidden layer and an output layer, and the input variables are weighted nonlinearly to finally obtain the output variables. The layers of the neural network interact and connect with each other and with each neuron on non-identical layers to form a complete system, and this system is characterized by self-adaptation, self-learning and information processing, in which the connection weights as well as the thresholds are in a dynamic process of change, while constantly outputting accurate predictions [12].

In this paper, the experimental hidden layer activation function f is a hyperbolic tangent function, which is given by:

$$tanh(x) = \frac{1 - exp(-2x)}{1 + exp(-2x)} \quad (2)$$

2.3.MLR-ARIMA model

MLR-ARIMA is a combined model analysis method that combines multiple linear regression (MLR) and autoregressive integrated moving average (ARIMA) models.The ARIMA model is mainly used to fit the regression residuals generated by the multiple linear regression model and make short-term predictions, and then the predicted residuals are overlaid with the predicted values of the multiple linear regression to obtain the predicted values of the combined model.

The general form of the ARIMA (p,d,q) model is:

$$u_t = \alpha + \varphi_1 u_{t-1} + \ldots + \varphi_t u_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_t \varepsilon_{t-q} \quad (3)$$

Where p is the autoregressive order; d is the number of differentials; q is the sliding average order; u_t is the differenced smooth series; α is a constant; φ is the autoregressive model coefficient; θ is the moving average model coefficient; ε_t is a zero-mean white noise sequence.

MLR-ARIMA model is calculated as:

$$y_{MLR-ARIMA} = y_{MLR} + u_{MLR\text{-}RES} \quad (4)$$

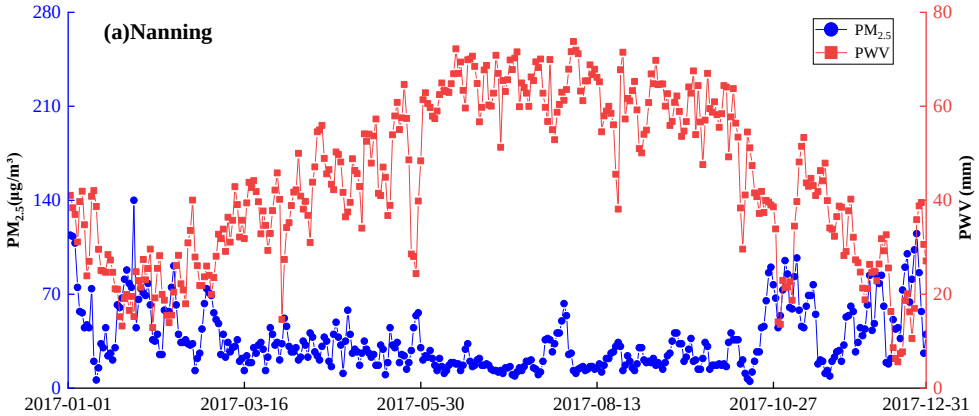
Where y_{MLR} is the predicted value of multiple linear regression; $u_{MLR\text{-}RES}$ is the predicted values for the residuals of the multiple linear regression model fitted for the ARIMA model。

2.4.Experimental data

2.4.1.Data source

The data used in this paper are the daily average concentrations of SO_2, NO_2, CO, O_3, and $PM_{2.5}$ for Nanning, Guilin, and Baise for a total of 365 days in 2017,(data from http://envi.ckcest.cn/environment/). In addition, the daily average of Precipitable Water Vapor (PWV) obtained from three sounding stations, 59431 Nanning, 57957 Guilin and 59211 Baise, for 365 days in 2017 was added(data from http://weather.uwyo.edu/upperair/sounding.html).

The changes in $PM_{2.5}$ concentrations as well as the changes in PWV values in the three Guangxi cities in 2017 were plotted as line graphs, respectively, and the results are shown in Figure 1.



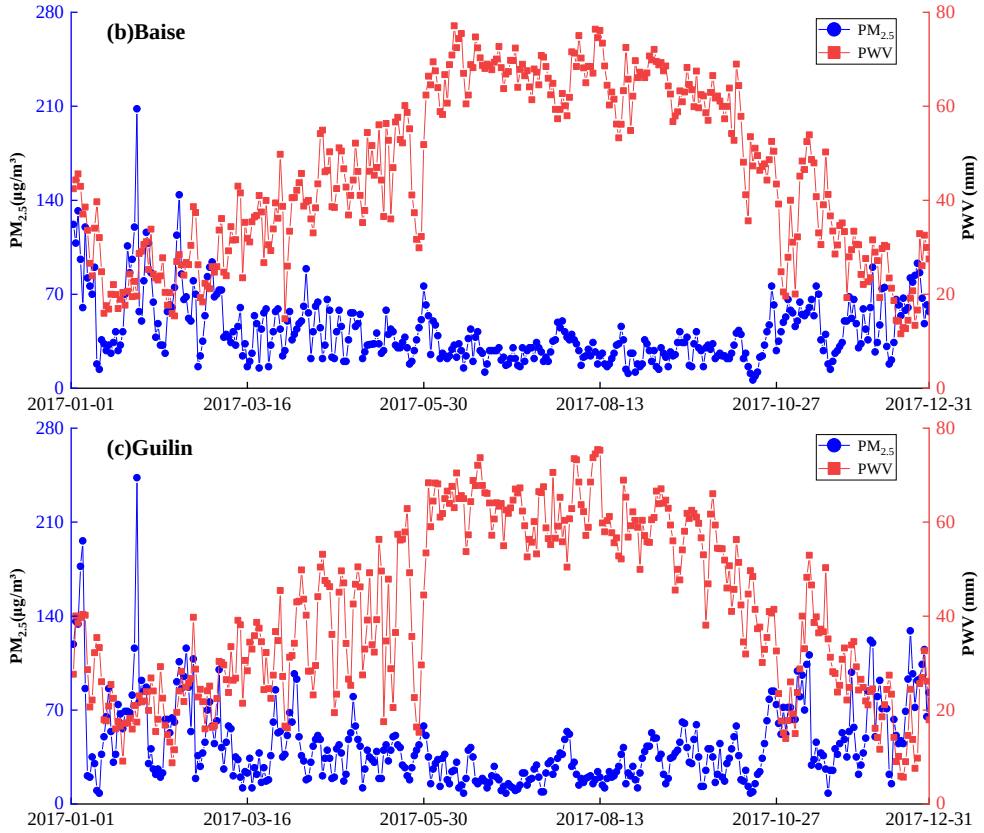


Figure 1 Changes in $PM_{2.5}$ and PWV values of three cities in Guangxi in 2017

From the Figure 1, we can see that the annual $PM_{2.5}$ concentration changes in major cities in Guangxi area show the characteristics of high at both ends and low in the middle, that is, the air quality is worse in winter and better in summer, possessing obvious seasonal characteristics; while the annual changes of PWV in Guangxi area show the opposite trend to $PM_{2.5}$, with the characteristics of low at both ends and high in the middle, especially between May and November. The difference between $PM_{2.5}$ and PWV changes is especially obvious, which is directly related to the more rain in summer and autumn in Guangxi area.

Analyzed from the perspective of individual cities, Guilin has the largest variation in $PM_{2.5}$ concentration, with its annual $PM_{2.5}$ peak close to 250 μg/m³, which is much higher than that of Nanning and Baise, followed by Baise, whose $PM_{2.5}$ peak can reach 200 μg/m³, while Nanning has the best air quality performance, with its annual peak not exceeding 150 μg/m³, which shows that although the Guangxi region has a better air quality performance in the This shows that despite the better air quality performance and developed tourism industry in the whole country, there is still a need to pay attention to the changes of $PM_{2.5}$, an air pollutant.

2.4.2.Correlation Analysis

Correlation analysis of the variables used in the modeling.

Table 1 Correlation analysis results of $PM_{2.5}$ and various variables in three cities in 2017

City	SO_2	NO_2	CO	O_3	PWV
Nanning	0.815	0.740	0.610	0.397	-0.573
Guilin	0.733	0.715	0.617	0.252	-0.445
Baise	0.550	0.642	0.312	0.184	-0.460

From Table 1, it can be seen that SO_2, NO_2, CO and O_3 are positively correlated with $PM_{2.5}$ and PWV is negatively correlated in Nanning, Guilin and Baise in 2017, which coincides with the

variation pattern in Figure 1, and their correlation levels are high, indicating that this factor can be considered for $PM_{2.5}$ concentration prediction, where the variable correlation levels in Nanning and Guilin are in the same high and low positions, in the following order. SO_2>NO_2>CO>PWV>O_3, while the correlation rank of variables in Baise City is in the order of SO_2>NO_2>PWV>CO>O_3. The correlation of variables in the table is significant at the confidence level (double test) of 0.01, which can prove that the various variables selected for modeling are well correlated with $PM_{2.5}$ and can be used for modeling prediction.

3.Results & Discussion

The SO_2, NO_2, CO, O_3, PWV, and $PM_{2.5}$ data of Nanning, Guilin, and Baise cities for the first 362 days of 2017 were used as the training set to establish the multiple linear regression model and the data of the last three days of 2017 were used as the validation set to verify the model prediction effect, and the model parameters are shown in the following table 2.

Table 2 Three-city multiple linear regression model parameters

City	R^2
Nanning	0.778
Guilin	0.729
Baise	0.559

In the table, R^2 is the goodness of fit of the model, and its value ranges from 0 to 1. The closer the value is to 1, the better the model fitting effect is, and from the data in Table 2, it can be seen that the best fitting effect of the multiple linear regression model is Nanning, followed by Guilin, and the poor fitting effect of Baise. , therefore, the predicted values obtained from the model and the corresponding Mean Absolute Error (MAE) results are calculated and included in Table 3, and the formula for calculating the MAE is as follows.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | X_i - \hat{X}_i | \tag{5}$$

Table 3 Three-city $PM_{2.5}$ multiple linear regression forecast results

City	Date	Actual value ($\mu g/m^3$)	Predicted value ($\mu g/m^3$)	Residual error ($\mu g/m^3$)	MAE($\mu g/m^3$)
Nanning	2017.12.29	57.00	50.73	6.27	13.41
	2017.12.30	26.00	45.54	-19.54	
	2017.12.31	40.00	54.41	-14.41	
Guilin	2017.12.29	115.00	87.23	27.77	14.88
	2017.12.30	65.00	48.57	16.43	
	2017.12.31	83.00	83.45	-0.45	
Baise	2017.12.29	48.00	68.36	-20.36	9.76
	2017.12.30	62.00	56.73	5.27	
	2017.12.31	57.00	60.65	-3.65	

The mean absolute error is calculated by computing the mean of the absolute values of the deviations of all observations in the sample from the corresponding estimated values. This calculation method can avoid the problem of sample errors canceling each other out, and thus can accurately reflect the magnitude of the actual prediction error.

As can be seen from Table 3, the multiple linear regression prediction results have two-day data residual values greater than 10 for both Nanning and Guilin, and one-day data residual values greater than 10 for Baise, while daily residual values over 20 exist for both Guilin and Baise; as can be seen from the mean value of the absolute residuals, the values for both Nanning and Guilin are above 10,

with Guilin's value already close to 15, while Baise Although it is below 10, it does not show a better performance. Therefore, the results in the table can prove that the prediction accuracy of the multiple linear regression model in the major cities of Guangxi region has a large difference, and its model prediction ability needs to be further improved.

Construct the SO_2, NO_2, CO, O_3, PWV, and $PM_{2.5}$ data of Nanning, Guilin and Baise for the first 362 days in 2017 as the training samples. The model parameters are shown in Table 4:：

Table 4 Three-city neural network model parameters

City	Training error rate	Test the error rate
Nanning	0.167	0.172
Guilin	0.177	0.159
Baise	0.471	0.319

From the perspective of parameters in Table 4, the neural network model, like multiple linear regression model, is better in Nanning and Guilin than Baise in training, and the model prediction results are shown in Table 5:

Table 5 $PM_{2.5}$ neural network prediction results in three cities

City	Date	Actual value ($\mu g/m^3$)	Predicted value ($\mu g/m^3$)	Residual error ($\mu g/m^3$)	MAE($\mu g/m^3$)
Nanning	2017.12.29	57.00	57.44	-0.44	14.15
	2017.12.30	26.00	48.57	-22.57	
	2017.12.31	40.00	59.43	-19.43	
Guilin	2017.12.29	115.00	114.37	0.63	11.24
	2017.12.30	65.00	53.49	11.51	
	2017.12.31	83.00	104.58	-21.58	
Baise	2017.12.29	48.00	70.24	-22.24	10.41
	2017.12.30	62.00	59.37	2.63	
	2017.12.31	57.00	63.36	-6.36	

From the prediction results in Table 5, Nanning and Guilin cities with better training of neural network model parameters showed good prediction effects on the first day of the prediction set, and their residual values were less than 1. However, the prediction effects on the second and third days were not good, but Baise city had better prediction results than Nanning and Guilin city on the second two days; and from the mean absolute error, it can be seen that the neural network model effects were not much different from the multiple linear regression model effects, in which the multiple linear regression model effects in Nanning and Baise city were due to the neural network model, while in Guilin city, the neural network model prediction effects were better than the multiple linear regression model.

To further improve the prediction accuracy of the multiple linear regression model, the ARIMA model is used to fit and analyze the regression residuals obtained from the regression of the training set of the multiple linear regression model.Since the ARIMA model has the advantage of time series fitting considering multiple regression, the modeling can be done with reference to the variation of other variables to determine its three main model parameters p, d, q.

The ARIMA model was used to forecast the residual values for the last three days of 2017 in the three cities, and the MLR-ARIMA model forecasts were obtained by overlaying the forecasts with the multiple linear regression model forecasts, and the results are shown in Table 6.

Table 6 $PM_{2.5}$ forecast results of MLR-ARIMA model

City	Date	Actual value ($\mu g/m^3$)	Predicted value($\mu g/m^3$)	Residual error ($\mu g/m^3$)	MAE($\mu g/m^3$)
Nanning	2017.12.29	57.00	51.25	5.75	7.57
	2017.12.30	26.00	41.84	-15.84	
	2017.12.31	40.00	41.13	-1.13	
Guilin	2017.12.29	115.00	93.30	21.70	12.75

	2017.12.30	65.00	50.04	14.96	
	2017.12.31	83.00	81.42	1.58	
	2017.12.29	48.00	69.50	-21.5	
Baise	2017.12.30	62.00	61.28	0.72	7.67
	2017.12.31	57.00	57.78	-0.78	

From the data in Table 6, we can see that the absolute value of the residuals of the MLR-ARIMA model in the three cities is reduced compared with the prediction results of the multiple linear regression model, in which the absolute value of the daily residuals in Nanning is greater than 10 on only one day, the absolute value of the residuals in Guilin is reduced on all three days, and the absolute value of the residuals in Baise is less than 1 on two days under the prediction of the MLR-ARIMA model. The prediction ability of the MLR-ARIMA model in Nanning City and Baise City is significantly improved compared with the other two types of models, and the prediction ability in Guilin City is better than that of the multiple linear regression model and slightly inferior to that of the neural network model.

The MAE for the prediction results of each model in the three cities is summarized in Table 7：

Table 7 Comparison of MAE results in three cities

City	MLR($\mu g/m^3$)	MLR-ARIMA($\mu g/m^3$)	Neural network($\mu g/m^3$)
Nanning	13.41	7.57	14.15
Guilin	14.88	12.75	11.24
Baise	9.76	7.67	10.41

From the summary results of MAE in Table 7, we can see that the MLR-ARIMA model improves the prediction results in Nanning by 43.55% and 46.50% compared to the multiple linear regression model and the neural network model, respectively, and in Baise by 21.41% and 26.32%, respectively, while the best model prediction in Guilin is the neural network model, which improves the prediction results compared to the MLR and MLR-ARIMA model by 24.46% and 11.84%, respectively, although the MLR-ARIMA model still has an accuracy improvement of 14.31% compared to the MLR model, so it is effective and feasible to perform quadratic fitting for the residuals of the multiple linear regression model.

In terms of model enhancement effects, Nanning has the best applicability to the MLR-ARIMA model, followed by Baise and slightly less effective in Guilin; while the neural network model shows better applicability in Guilin, especially on the first day of forecasting.

4.Conclusions

In this paper, a three-day short-term prediction of $PM_{2.5}$ concentrations was conducted for three major prefecture-level cities in the Guangxi region by combining their urban air pollutant concentration change data in 2017 and the atmospheric precipitable water PWV data obtained from sounding stations located in the three cities, and the following conclusions can be drawn.

(1) The changes of PWV values in Guangxi region show the characteristics of low at both ends and high in the middle, while the changes of $PM_{2.5}$ concentration show the characteristics of low in the middle and high at both ends, both showing obvious seasonal characteristics.

(2) The PWV data has a high correlation grade with PM2.5, and shows negative correlation with PM2.5 in the process of annual change, so the comprehensive consideration of PWV influence can be used in the analysis and prediction of the causes of $PM_{2.5}$ in Guangxi area.

(3) The applicability and prediction accuracy of MLR-ARIMA model in three cities of Guangxi are better than the multiple linear regression model, among which Nanning has the most obvious improvement of accuracy and the best applicability of the model, while the applicability of the model in Guilin and Baise have different degrees of improvement compared with the multiple linear regression model; compared with the neural network model, it shows the better effect in Nanning and

Baise and weaker in Guilin. However, the overall effect is still good, so the model can be used as a reference for $PM_{2.5}$ prediction in major cities in Guangxi region.

Acknowledgments
This paper was one of the phase results of the National Natural Science Foundation of China Regional Science Foundation Project "Research on Guangxi model of smog-haze forecasting based on geographically weighted regression Kriging" (41864002).

References
[1] Yan Z N, Mou J F, Zhao X, et al. Time series prediction analysis of atmospheric $PM_{2.5}$ concentration in Shenzhen based on ARIMA model[J]. Modern preventive medicine, 2018, 45(02): 220-223+242.
[2] Asha B. Chelani. Estimating $PM_{2.5}$ concentration from satellite derived aerosol optical depth and meteorological variables using a combination model[J]. Atmospheric Pollution Research, 2019, 10: 847-857.
[3] Doreswamy, Harishkumar K S, Yogesh KM, et al. Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models[J]. Procedia Computer Science, 2020, 171: 2057–2066.
[4] Xie S F, Zhao Y, Li G H, et al. PM2.5 concentration prediction based on WPA-WOA-BP neural network[J]. Journal of Geodesy and Geodynamics, 2021, 41(01): 12-16.
[5] LI J X, LIU X S, LIU J, et al. $PM_{2.5}$ concentration prediction based on MRMR-HK-SVM model[J]. China Environmental Science, 2019, 39(06): 2304-2310.
[6] Zhang S C, Li Z Y, Dai K Y, et al. Correlation of GPS water vapor variation and haze in Beijing[J]. Science of Surveying and Mapping, 2016, 41(08): 43-47.
[7] Zhou Y J, Yao Y B, Xiong Y L, et al. Correlation study of PWV and $PM_{2.5}$ based on Spearman's rank correlation coefficient[J]. Journal of Geodesy and Geodynamics, 2020, 40(03): 236-241.
[8] Guo T J, Yao Y B, Zhou Y J. A random forest prediction model for $PM_{2.5}$ by fusing GNSS meteorological parameters[J]. Science of Surveying and Mapping, 2021, 46(04): 37-42+56.
[9] Wang Y, Liu Y P, Li J B, et al. Effects of water vapor and wind speed on $PM_{2.5}$/PM_{10} variation in haze[J]. Journal of Catastrophology, 2015, 30 (1) : 5-7.
[10]Wang Y, Liu Y P, Li J B, et al. A comparative study of GPS water vapor and $PM_{2.5}$ mass concentration in Hebei Province[J]. Journal of Geodesy and Geodynamics, 2016, 36(01): 40-42.
[11]Wang Y, Ren D, Liu Y P, et al. Modeling of spring $PM_{2.5}$ concentration in Hebei Province by integrating GNSS PWV, wind speed and air pollution observation[J]. Geomatics and Information Science of Wuhan University, 2019, 44(08):1198-1204.
[12]Tan W H. Research on logistics demand forecasting in Jiangxi Province based on multiple regression and neural network [D]. Nanchang:Jiangxi University of Finance and Economics, 2020.