The research of PM2.5 concentrations model based on regression calculation model

Cite as: AIP Conference Proceedings **1794**, 030005 (2017); https://doi.org/10.1063/1.4971927
Published Online: 06 January 2017

Junmin Li, and Luping Wang

The Research of PM2.5 Concentrations Model Based on Regression Calculation Model

Junmin Li, Luping Wang [a)]

College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an, 710054, China.

[a)]Corresponding author: 18706703126@163.com

Abstract. In this paper, we mainly use the urban air quality monitoring data as the study data, the relationship between PM2.5 concentrations and several major air pollutant concentrations, meteorological elements in the same period are analyzed respectively. Through the analysis, we find that there is a significant correlation characteristic between the concentrations of PM2.5 and the concentrations of PM10, SO_2, NO_2, O_3, CO, temperature and humidity. So, we take these factors as the variables; make a multiple linear regression analysis about PM2.5 concentrations, set up the urban PM2.5 concentrations regression calculation model. Through the estimation results of the model. In comparisons with the observed results, two results are basically identical, which shows that the regression model has a good fitting effect and use value.

Key words: The concentration of PM2.5; Multiple linear regression; Correlation coefficient; Meteorological elements.

INTRODUCTION

At present, with the rapid development of our country's economy, the environment problem is increasingly serious. Urban pollution comes mainly from the waste discharged by industry, enterprise, life stoves, heating boiler, transportation and so on, most of them contain heavy metals and other hazardous substances, which not easily ejected by human body. However, The PM2.5 particle attracts people's widespread attention, which radius is minimal and inclines to carry toxic substances, makes seriously effect to human health. Therefore, the research of PM2.5 concentration in the atmosphere is of great significance for air quality assessment and take measures of air quality control.

The problem of PM2.5 not only relates to public life and urban governance, but also relates to the national image and social stability. The PM2.5 monitoring starts in 2012 in Our country. However, because of PM2.5 monitoring facilities are quite expensive, and each of them needs room to placement, which requires huge capital investment, and the cost is rather high [1, 2]. Therefore, the environment monitoring center sites in each large or medium-sized city very few at present, and the monitoring data can't reflect the city's whole air quality. In order to reflect the local air quality comprehensively and completely, we need to make analysis model. Which on the basis of monitoring data of other air pollutants and meteorological elements to analyze the concentrations of PM2.5 in the air. The simulation results and the monitoring results are complementary, which can meet the public demand about environment, improve the image of the government.

The formation of PM2.5 in urban atmosphere is complicated. On the one hand, under a certain environment condition, part of the atmosphere gaseous pollutants (such as SO_2, NO_2, CO, etc) can form PM2.5; On the other hand, the content of PM2.5 in different place in a city is significantly different during the same time. However, Existing researches generally only focus on pollutants variation trends over time, the researches for the relationship between concentrations of different pollutants and meteorological elements are scarcely ever. In this paper, we mainly make a correlation analysis about PM2.5 concentrations and several major air pollutants concentrations,

2016 International Conference on Materials Science, Resource and Environmental Engineering
AIP Conf. Proc. 1794, 030005-1–030005-7; doi: 10.1063/1.4971927
Published by AIP Publishing. 978-0-7354-1460-0/$30.00

meteorological elements in the same period, and constitute a regression calculation model about the concentration of PM2.5

RESEARCH DATA

This paper, we mainly take xi'an as an example, carry out the research about regression calculation model about the concentration of PM2.5. The research data is the monitoring data of xi 'an thirteen air quality monitoring sites during the past two years. Where we can gather the concentrations data of PM2.5, PM10, SO_2, NO_2, O_3, CO, temperature and humidity data and so on. The monitoring data recorded continuously by automatic recording instrument for 24 hours a day, one datum is got in an hour. In this study, we take the mean value of unification xi'an 13 sites monitoring data as the research data, constitute and test the regression calculation model about the concentration of PM2.5.

First, we should make a arithmetic average about the 24-hour monitoring data of xi'an 13 air quality monitoring sites to get the unification daily mean concentration data of PM2.5, other gaseous pollutants, temperature and humidity; Then build xi'an unification concentration calculation model of PM2.5, make the reliability verity for this model .

Experiment designed as follows: first,we unified analysis the monitoring data from 13 monitoring sites in xi'an, get unified analysis data,monitoring site data update one time a hour. In this paper, we take the mean value of Unification xi'an 13 sites monitoring data as the research data, including, 30 days' data extracted randomly for reliability test, and the rest data are used to structure the model of PM2.5 concentrations. As shown in figure 1 is the statistical analysis result of xi'an air quality and the pollutant concentration during April 2015 to April 2016, the graph shows that there is a certain relationship between the concentration of PM2.5 and the other observation elements, However we need to establish a model for quantitative analysis to determine what the specific relationship is.

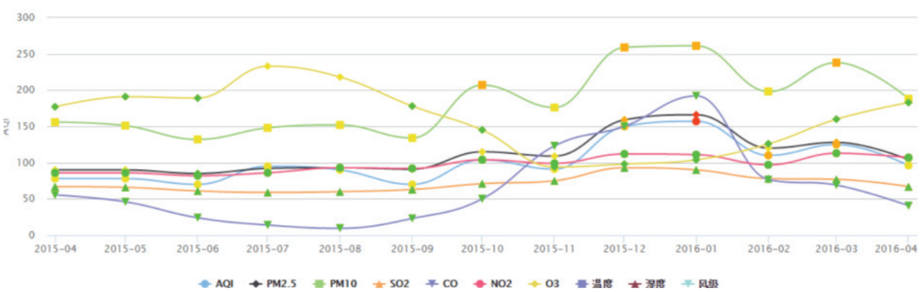


FIG. 1 The air quality graph of Xi'an (2015.4-2016.4)

RELEVANT FEATURE ANALYSIS BETWEEN PM2.5 AND OTHER OBSERVED ELEMENTS

In order to constitute the regression calculation model about the concentrations of PM2.5, we need to know that which factor has strong correlation with PM2.5. Therefore, the have to make correlation analysis between PM2.5 concentrations and other observation elements severally.

The double variable correlation indexes calculate formula:

$$r=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{\sum(Y-\overline{Y})^2}}=\frac{\sum XY-\frac{(\sum X)(\sum Y)}{n}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{n})(\sum Y^2-\frac{(\sum Y)^2}{n})}} \tag{1}$$

X and Y are two variables, is the sample size.

Then, we can get the bivariate correlation analysis result as follow.

TABLE1. The results of bivariate correlation analysis between pm2.5 and relevant factors

Variables	Correlation coefficients	Significance level	Sample size
PM2.5 & PM10	0.908	0.000	732
PM2.5 & SO2	0.762	0.000	732
PM2.5 & NO2	0.663	0.000	732
PM2.5 & CO	0.834	0.000	732
PM2.5 & O3	-0.405	0.000	732
PM2.5 & Temperature	-0.457	0.000	732
PM2.5 & humidity	0.015	0.775	732

We can see from Table 1, The Significance level between PM2.5 and PM10, SO_2, NO_2, O_3, CO, temperature is 0, less than 0.05, it shows that they have pass the significance test, and can be used for building PM2.5 concentration calculation model. However the significance level between PM2. 5 and humidity is 0.775 (> 0.05), failed the significance test.

It is a significant positive correlation between PM2.5 and PM10, SO_2, NO_2, CO, and it is a significant negative correlation between PM2.5 and O_3, temperature.

MODEL CONSTRUCTION

Construct PM2.5 Concentrations Model Based On Multivariate Linear Regression

Because of the rest of bivariate correlation analysis between PM2.5 and PM10, SO_2, NO_2, O_3, CO, temperature pass significance test, therefore, we use the monitoring data of PM10, SO_2, NO_2, O_3, CO and temperature to build the PM2.5 concentrations model. The dependent variable is the concentration of PM2.5 (Y), the independent variables are PM10 (X_1), SO_2 (X_2), NO_2 (X_3), CO (X_4), O_3 (X_5) and temperature (X_6).

Let m be the number of independent variables, n be the number of samples, we can set up multiple linear regression forecast model as follows:

$$\begin{cases} y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} \cdots \beta_m X_{1m} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} \cdots \beta_m X_{2m} + \varepsilon_2 \\ \cdots\cdots \\ y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} \cdots \beta_m X_{nm} + \varepsilon_n \end{cases} \tag{2}$$

Let β_0 be the constant, ε be the observation error: $\varepsilon_i \sim N(0,\sigma^2)$ ($i = 1 \cdots n$), and $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ are independent to each other.

Let

$$\underset{n\times 1}{Y} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix}, \quad \underset{n\times(m+1)}{X} = \begin{Bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{Bmatrix}$$

$$\underset{(m+1)\times 1}{\beta} = \begin{Bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{Bmatrix}, \quad \underset{n\times 1}{\varepsilon} = \begin{Bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{Bmatrix}$$

We show (2) formula with a matrix:

$$\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, \mathrm{Cov}(\varepsilon, \varepsilon) = \sigma^2 I_n \end{cases} \tag{3}$$

We can use the least square method to get the unbiased estimate, effective and consistent estimate about the multiple linear regression model. And the Estimate formula as follows:

$$\hat{\beta} = \begin{Bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{Bmatrix} = (X^T X)^{-1} X^T Y \tag{4}$$

Then, we can get the estimates of $\beta_0, \beta_1, \cdots, \beta_m$ are $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_m$, where

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_m) &= \min_{\beta_0, \beta_1, \cdots, \beta_m} Q(\beta_0, \beta_1, \cdots, \beta_m) \\ &= \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i \cdots - \beta_m x_i)^2 \end{aligned} \tag{5}$$

We can get the regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \cdots \hat{\beta}_m x_m \tag{6}$$

Let $m = 6, n = 732$, the departments of the model calculated by SPSS software shown in table 2.

TABLE 2. The multiple linear regression coefficient table between pm2.5 and relevant factors

model	non-standard coefficients		Trial standard coefficients	t	Sig.
	B	standard error			
constant	-32.165	3.782		-8.505	.000
X_1	.348	.016	.622	21.804	.000
X_2	.594	.113	.189	5.241	.000
X_3	-.227	.065	-.092	3.500	.001
X_4	20.247	1.804	.375	11.222	.000
X_5	.032	.028	.034	1.176	.024
X_6	.504	.156	.111	3.242	.001

We can get from table 2, the probability value of regression coefficient in the significance test, all the significance level Sig are less than 0.05, so the model is feasible. The result is:

$$\begin{aligned} Y &= -32.165 + 0.348 X_1 + 0.594 X_2 - 0.227 X_3 \\ &+ 20.247 X_4 + 0.032 X_5 + 0.504 X_6 \end{aligned} \tag{7}$$

The unit of PM2.5(Y), PM10(X_1), SO_2(X_2), NO_2(X_3) and O_3(X_5) is $\mu g / m^3$, the unit of CO (X_4)is mg / m^3 ,The unit of temperature (X_6) is Celsius (°C).

Reliability Test

Do reliability test with the random extraction Xi’an 30 days’ monitoring data in advance? The comparisons of calculate results and measured results are showed in figure 2. We can see that, the imaginary line and the solid line are unanimous, shows that the result of PM2.5 concentration regression calculation model and the measured results are unanimous, and the presented model in this paper can achieve a good fitting effect in Xi’an PM2.5 concentration calculation.

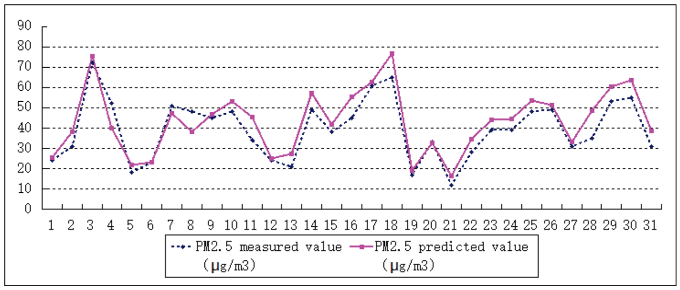
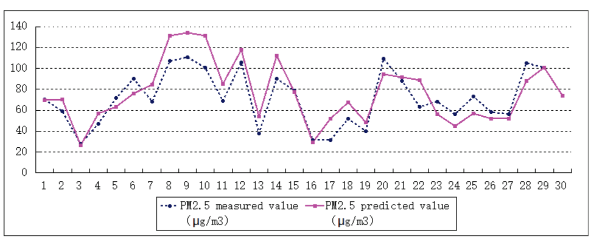
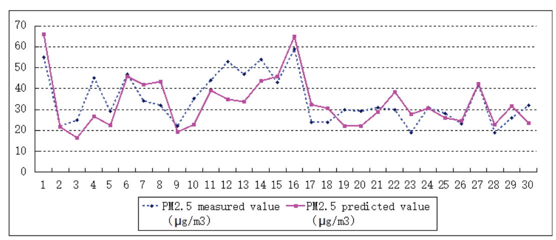


FIG. 2 The comparisons results of model prediction results and measured results of Xi 'an

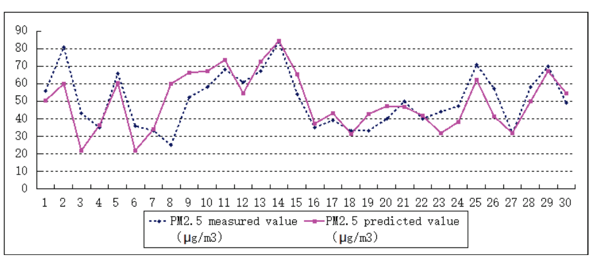
In addition, in order to verify whether the model can also be applied to other cities or regions, we randomly drawing four cities from north China, central China, east China and northeast. Randomly collecting 30 days monitoring data of PM2.5 density and other related elements, to estimate the result with the PM2.5 model presented in this paper, and paint into the line chart with the measured values and the estimated ones, as shown in figure 3, the model we conceive also have better good effect for the four different urban areas, shows that this calculation model of PM2.5 concentrations is generality for other cities.



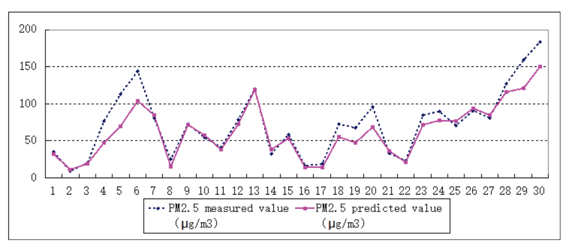
(a) Zhengzhou



(b) Changchun



(c)Hefei



(d) Beijing

FIG. 3 The comparisons results of model prediction results and measured results

As we can see through the analysis above: it is a complex process that the formation and the content changing in the air. And in the process, the relationship between PM2.5 concentrations and the concentrations of PM10, SO_2, NO_2, O_3, and CO, temperature has an obvious linear characteristic. So, it is reasonable and effective to use multiple linear regression equations to reflect the dependencies between air pollutant concentrations and other meteorological elements. As shown from the model test results, the PM2.5 concentrations model presented in this paper is effective. It shows that is possible to use the monitoring concentrations of PM10, SO_2, NO_2, O_3, CO and temperature to calculate PM2.5 concentrations. Which can be used to calculate PM2.5 concentrations in the areas where do not have PM2.5 monitoring conditions.

CONCLUSION

In this paper, we use the recent two years monitoring data of Xi'an 13 air quality monitoring sites as experimental data, and make the double variant correlation analyses between PM2.5 concentrations and other monitoring elements severally, the result shows that the concentrations of PM2.5 and PM10, SO_2, NO_2, O_3, CO and temperature usually have obvious linear relationship during this same period. At the same time, the production, formation and changing process of different pollutants are correlated, the formation and change of process exists between interconnected. Therefore, we construct a multiple linear regression calculation model between the concentrations of PM2.5 and PM10, SO_2, NO_2, O_3, CO and temperature. And make a model test with reserved 30 days monitoring data. Meanwhile, in order to verify whether the model is universal, we make a reliability test for the model with respective 30 days monitoring data from four different cities randomly, the main conclusions are as follows:

(1) First, they are significant positive relevant between the concentration of PM2.5 and PM10, SO_2, NO_2 and CO, they are significant negative relevant between the concentration of PM2.5 and O_3, temperature, and all these elements can be used for building PM2.5 concentration calculation model. We can see from the values of absolute value of correlation coefficients, the correlation with PM2.5 from large to small are PM10, CO, SO_2, NO_2, temperature and O_3.

(2) Then, we have made a reliability test about the model with reserved Xi'an 30 days monitoring data. Result shows that the PM2.5 concentration between the model calculation and the actual monitoring are basically consistent, imitative effect is very good, it shows that this model can be used to forecast the PM2.5 concentrations of Xi'an.

(3) Last, we selected four cities from different parts of our country randomly, and collect 30 days monitoring data of PM2.5 density and other related elements from each city randomly, to estimate the result with the PM2.5 model presented in this paper. The result shows that this model is effective for the four different cities. Therefore, this PM2.5 concentrations regression calculation model is widely used for other cities.

REFERENCES

1. Dai Yongli, TaoJun, Lin Zejian, et al. China's megacities haze weather characteristics and the influence factor analysis during 2006-2009. Journal of environmental science, 2013 (8): 2925-2932.
2. Meng Xiaoyan, Rui-bin Wang, Zhang Xin, et al. The main pollutant concentration changes of key environmental protection cities during 2006-2010. Journal of environmental science research, 2012, 25 (6): 622-627.
3. Hara K, Homma J, Tamura K, et al. decreasing trends of suspended particulate matter and PM2. 5 concentrations in Tokyo, 1990-2010. Journal of the Air & Waste Management Association, 2013, 63(6): 737-748.
4. Bigi A, Ghermandi G. Long-term trend and variability of atmospheric PM10 concentration in the Po Valley. Atmospheric Chemistry and Physics, 2014, 14(10):4895-4907.
5. Anne Boynard, Cathy Clerbaux, Lieven Clarisse, et al. First simultaneous space measurements of atmospheric pollutants in the boundary layer from IASI: A case study in the North China Plain. Geophysical Research Letters, 2014, 41(2): 645-651.
6. Zhao X J, Zhao P S, Xu J, et al. Analysis of a winter regional haze event and its formation mechanism in the North China Plain. Atmospheric Chemistry and Physics, 2013, 13:5685- 5696.
7. Tao Jinhua, Zhang Mei-gen, Chen Liang-fu, etc. A new method of particle concentration near the ground based on satellite remote sensing AOT estimation. Chinese science (D), 2013 lancet (1): 143-154.
8. Levy R C, Remer LA, Mattoo S, et al. Second-generation operational algorithm: Retrieval of aerosol properties over land from in-version of Moderate Resolution Imaging Spectro radiometer spectral reflectance. Journal of Geophysical Research, 2007, 112, D13211:1-21.
9. Zhang Ju, Miao Hong, Ouyang Zhi-yun, etc. The analysis of Beijing suburbs environment air quality change and its impact factor nearly 20 years. Journal of environmental science, 2006 (11): 1886-1892.

10. Wang Hua, Jiang Nan, Yang Naiwang. The study of atmospheric environment quality and its change trend in Xi 'an during 1991-2012. Journal of environmental engineering, 2014 (suppl 1): 526-529.