# *Environmental governance based on multiple linear regression fitting*

***Yu Zou[1*], Jinrun Xu[2]***

***((1. Southeast University Chengxian College, Jiangsu, China)***

***(2. Huanggang Normal University, Hubei, China)***

**ABSTRACT**

Regarding the construction of a mathematical model between air quality index (AQI) and different pollutant concentrations, we constructed pollutants respectively: PM2.5, PM10, CO, NO2, O3, SO2. firstly, we pre-processed the data in Annex 1, and analysed the data by visualising its missing values and boxplots to get no obvious outliers and missing values between the data. Secondly, we established multiple linear regression to fit the linearity between AQI and pollutant concentration, and fitted the expression: $y = 15.431 + 0.71 * X_1 + 0.077 * X_2 - 0.24 * X_3 + 11.867 * X_4 + 0.386 * X_5 + 0.273 * X_6$, with the goodness-of-fit of $R^2 > 90\%$ as good, and T-test and F-test were conducted to prove the significant difference. prove that there are significant differences. Secondly, considering the influence of multicollinearity between the constructed indicators, we introduced the variance inflation factor VIF to test the indicators, and concluded that the construction of indicators is reasonable. At the same time, we take the pollutants as the sub-sequence and the AQI index as the parent sequence to construct the grey correlation analysis, to derive the grey correlation between each of its pollutants and the AQI, and to rank them, see Table 3.In this paper, based on the constructed regression equations, the collected air quality data of the national cities are fitted, and it is concluded that the ten cities with the best air quality are Dazhou City, Hegang City, Heihe City, Jiamusi City, Shuangyashan, Yichun, Nanchong, Qiqihar, and Suining.

**Keywords:** Multiple Linear Regression; *VIF*; Grey Correlation Analysis; Entropy Weighting Method; TOPSIS; Simulated Annealing Algorithm, Minimum Number of Monitoring Points

## 1 INTRODUCTION

With the rapid development of the economy and the continuous growth of the population, China's environmental problems have become an important issue that needs to be solved urgently. These environmental problems have not only had a great impact on people's health and quality of life, but have also caused great damage to the ecosystem and ecological balance. In recent years, the Chinese government has actively promoted the development of environmental protection, and has achieved certain results by adopting a series of policies and measures, but the problem of environmental pollution is still very serious, and further measures need to be taken to solve it.

## 2 RELEATED WORK

Based on the information collected by our team, we built mathematical models between the Air Quality Index (AQI) and the concentration of different pollutants to better understand

the air quality situation and to give measures that can improve the air quality, rating the 10 best cities in the country in terms of air quality in each city.

## 3 MODEL ESTABLISHMENT AND SOLUTION

### *3.1 Problem 1 modeling*

#### 3.1.1 Data Preprocessing

Before establishing the relevant model solution, we first preprocess the data collected in our annex. One-dimensional interpolation of missing data values is carried out, and data outliers are eliminated to ensure that the data are from the correct source, and the missing data values visualised in this paper are as follows:

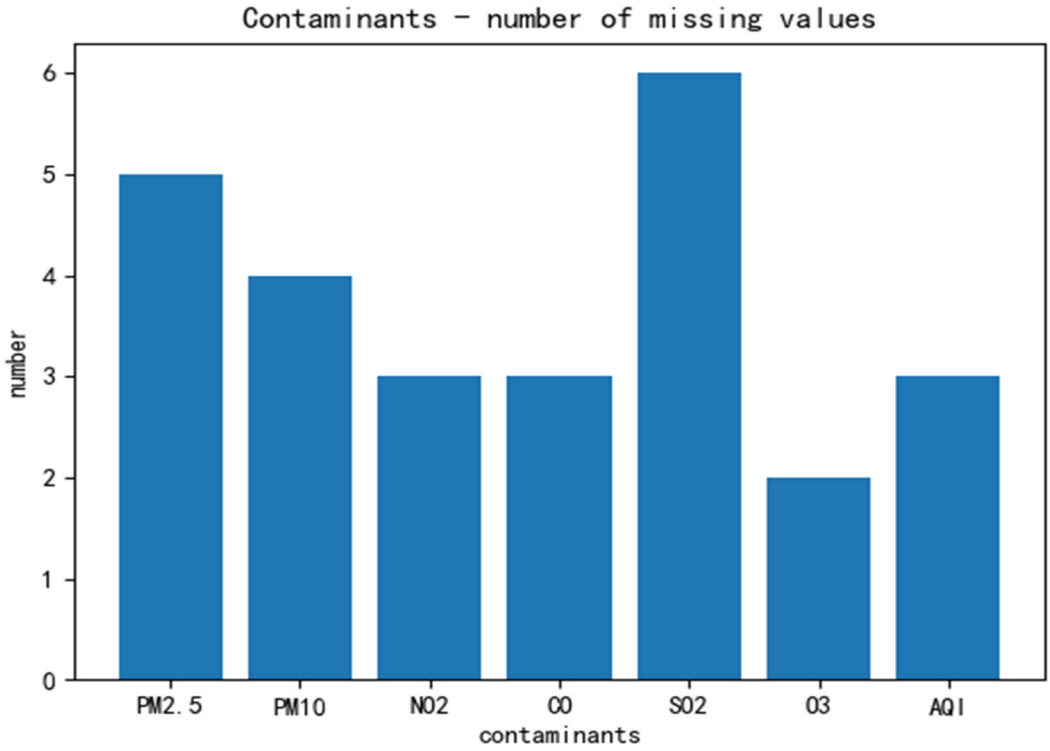


Figure 1: Pollutant data pre-processing

From Figure 1, we can see that there are missing values in the data indicators we collected, and the missing values of each indicator can be approximated within a reasonable range for interpolation.

Secondly after completing the visualisation of missing values, we perform residual analysis on the data to screen out the data outliers and remove them. The boxes in the box-and-line plot represent the middle 50% of the data, and the line inside the box represents the median." Whiskers" (tentacles) then extend beyond the box to represent the overall range of the data. Typically, the whiskers extend 1.5 times the interquartile range (i.e., the distance between the upper and lower quartiles) beyond the maximum and minimum values. In this paper, by visualising the box-and-line diagram, the outliers between the data of each indicator can be further analysed to provide relevant theoretical support for the establishment of the relevant

model later, and the visualised box-and-line diagram of this paper is as follows:

As can be seen in Figure 2, our overall collection of indicator data as a whole is consistent

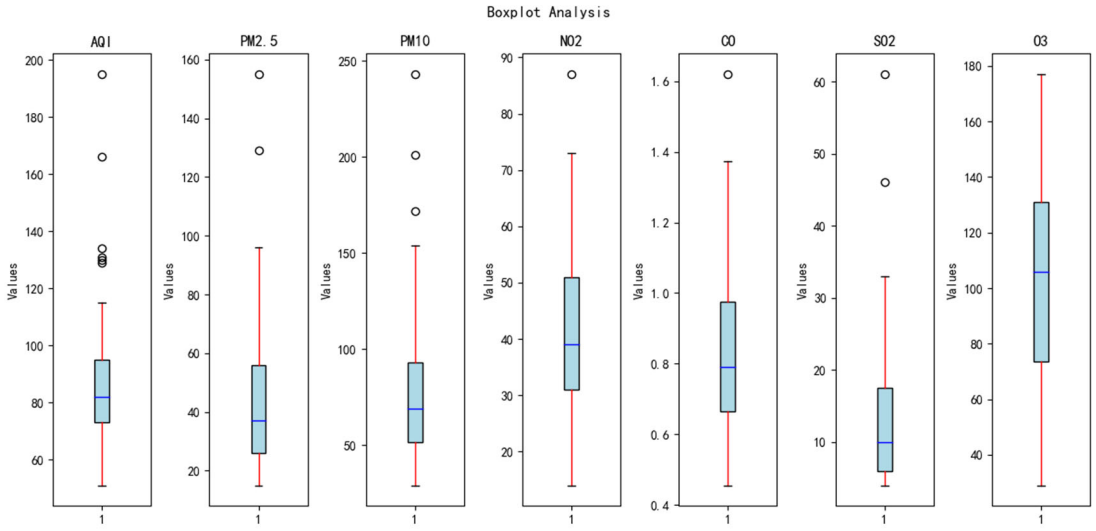


Figure 2: Data Error Box Plot

with the existence of a boxed interior i.e. within the middle 50% of the data, and there is no significant number of outlying data points, proving that there are no significant discrepancies in the data collected, which can be used to carry out the analysis of the model.

**3.1.2 Principle of Multiple Linear Regression Models**

Let the linear regression model of random y with $x_1, x_2, \ldots, x_k$ variables be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \quad (1)$$

Where is k+1 unknown parameters, called the regression constant, called the regression coefficient; y is called the explanatory variables; is k can be precisely controlled by the general variables, called the explanatory variables [1].

When p=1, the above equation is a univariate linear regression model, and when k≥2, the above equation is called a multivariate regression model. ε is the random error, which is the same as that of univariate linear regression, and is usually assumed to be the same as that of univariate linear regression [2].

$$\begin{cases} E(\varepsilon) = 0 \\ var(\varepsilon) = \sigma^2 \end{cases} \quad (2)$$

Similarly, the multivariate linear overall regression equation is.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad (3)$$

The coefficients represent the average units of the dependent variable y caused by a one-unit change in the independent variable, holding other independent variables constant [3]. The other regression coefficients have similar meanings, and in an aggregate sense, the multiple regression is a plane on a multidimensional space.

The multiple linear sample regression equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k \quad (4)$$

The estimation of the regression coefficients in the multiple linear regression equation can

also be done by the least squares method. From the residual sum of squares.

$$SSE = \sum(y - \hat{y}) = 0 \tag{5}$$

According to the principle of minimisation in calculus, the residual sum of squares SSE is known to have a minimal value [4]. In order to minimise the SSE, the partial derivatives of the SSE with respect to $\beta_0, \beta_1, \cdots \beta_k$must be zero.

The partial derivatives of SSE with respect to $\beta_0, \beta_1, \cdots \beta_k$, and make them equal to zero, can be obtained after finishing the k+1 equations:

$$\frac{\partial SSE}{\partial \beta_i} = -2\sum(y - \hat{y}) = 0 \tag{6}$$

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum(y - \hat{y})\, x_i = 0 \tag{7}$$

By solving this system of equations, the estimates of $\beta_0, \beta_1, \cdots \beta_k$ can be obtained respectively $\hat{\beta}_0$, $\hat{\beta}_1$, $\cdots\hat{\beta}_k$ regression coefficients of the estimated value of the independent variables when the number of independent variables is more, the calculation is very complex, must rely on the computer to complete independently. Now, with SPSS, the results are immediately available by simply entering the data and specifying the dependent variable and the corresponding independent variables [5].

For multiple linear regression, it is also necessary to determine the degree of fit of the equation and test the significance of the regression equation and regression coefficients.

Determining the degree of fit for multiple linear regression is similar to the coefficient of determination in univariate linear regression using the multiple coefficients of determination, where defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \overline{y})^2} \tag{8}$$

where SSR is the sum of squares of regression, SSE is the sum of squares of residuals, and SST is the sum of squares of total deviations.

Similar to the same linear regression $0 \le R^2 \le 1$ , the $R^2$ closer is to 1, the higher the fit of the regression plane, and vice versa, the closer$R^2$ is to 0, the lower the fit. The square root of $R^2$ becomes the negative correlation coefficient (R), which also becomes the multiple correlation coefficient [6]. It indicates the degree of linear correlation between the dependent variable y and all independent variables, which actually reflects the degree of correlation between the sample data and the predicted data. The magnitude of the coefficient of determination $R^2$ is affected by the number k of independent variables x. In the actual regression analysis, it can be seen that as the number of independent variables x increases, the sum of squares of regression (SSR) increases, is $R^2$ increases. Since the increase in $R^2$ caused by increasing the number of independent variables has nothing to do with you and good or bad, so when comparing the degree of fit between regression equations with different numbers of independent variables k, $R^2$ is not an appropriate indicator, and must be corrected or adjusted.

Adjustment method is: the residual sum of squares and the total deviation sum of squares of notes of the numerator denominator, respectively, divided by their respective degrees of

freedom, into the ratio of mean squared deviation, in order to remove the effect of the number of independent variables on the goodness of fit. The adjusted $\overline{R}^2$ is:

$$\overline{R}^2 = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}} = 1 - \frac{SSE}{SST} \bullet \frac{n-1}{n-k-1} = 1 - (1-R^2)\frac{n-1}{n-k-1} \tag{9}$$

As can be seen from the above, $\overline{R}^2$ takes into account the average residual sum of squares rather than the residual sum of squares, and therefore, generally in linear regression analyses, the larger $\overline{R}^2$ is the better [8].

The degree of fit of the regression equation can also be reflected from the F-statistic. A combined conversion of the formula for the F-statistic with the formula for $R^2$ gives:

It can be seen that if the fit of the regression equation is high, the more significant the F-statistic is; the F-statistic is significant in two months, the better the fit of the regression equation is:

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} \tag{10}$$

**3.1.3 Linear fitting of AQI and pollutants based on multiple linear regression**

Before establishing the pollutant-AQI regression fitting model, we first portrayed the scatter plot of the six major pollutants and analysed their trends. Secondly, we establish the linear fitting of AQI and pollutants based on multiple linear regression [9]. For the establishment of this model, we will further carry out the test of the sensitivity of the model solution, which corresponds to the introduction of the F-test, t-test, and the fitting of

the $R^2$ in order to assess the goodness of the establishment of the model.

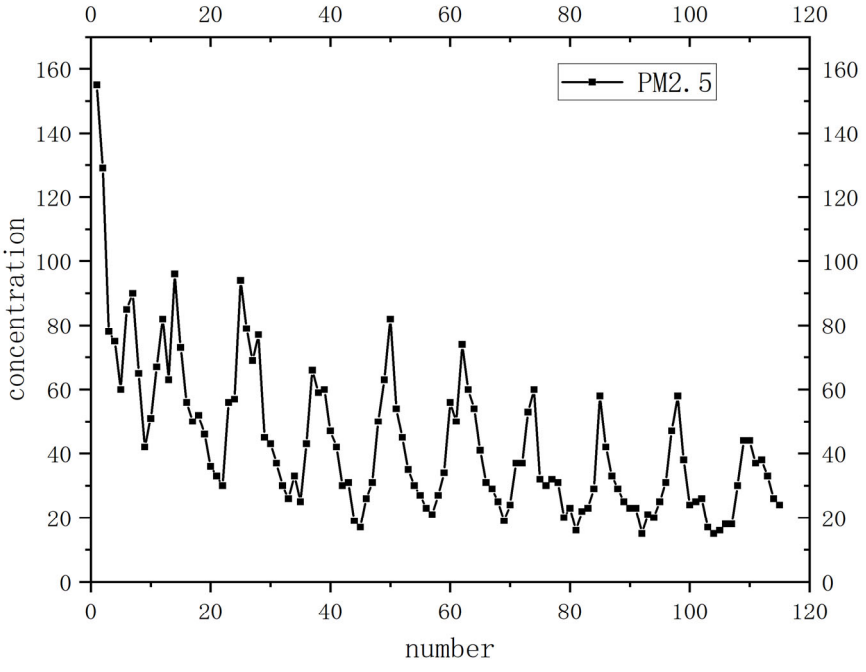

Figure 3:PM2.5 Concentration change chart

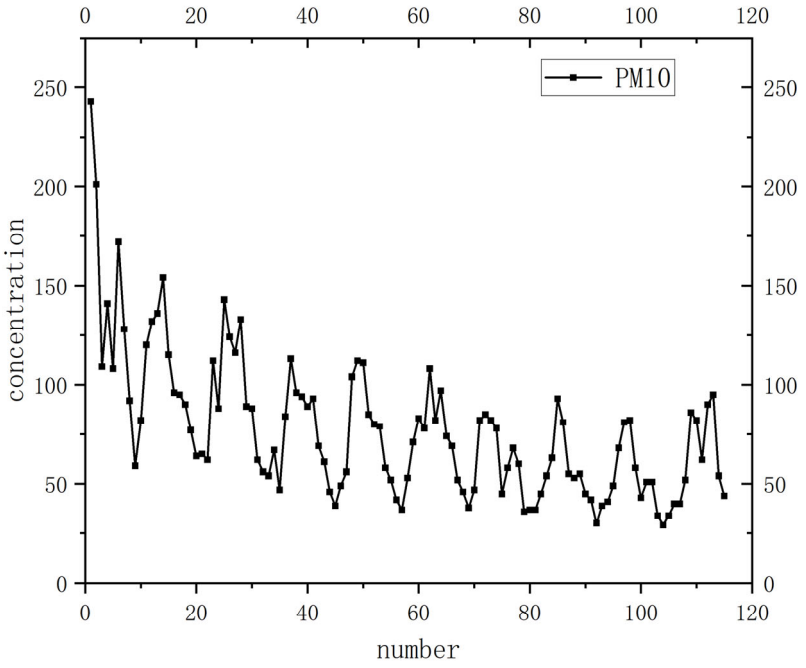


Figure 4:PM10 Concentration change chart

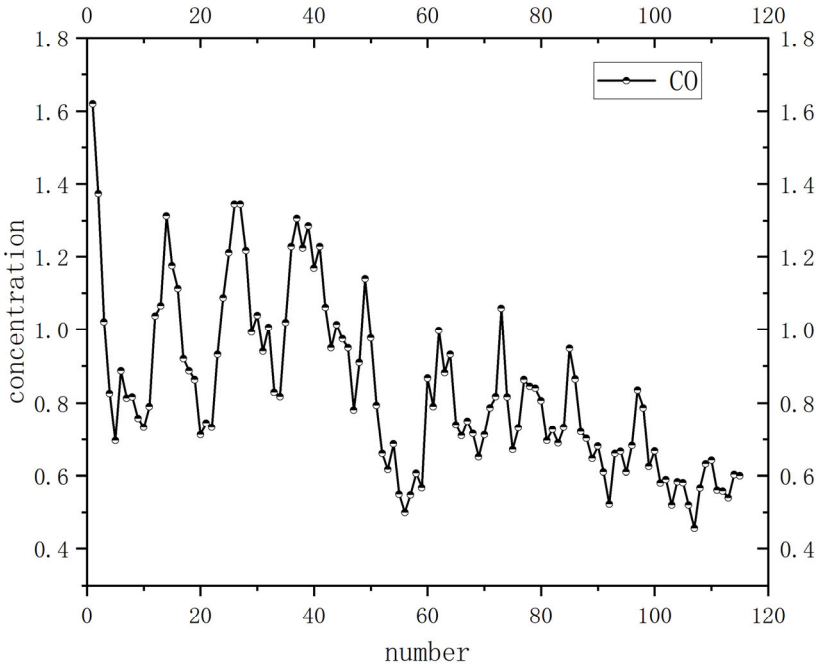


Figure 5:CO Concentration change chart

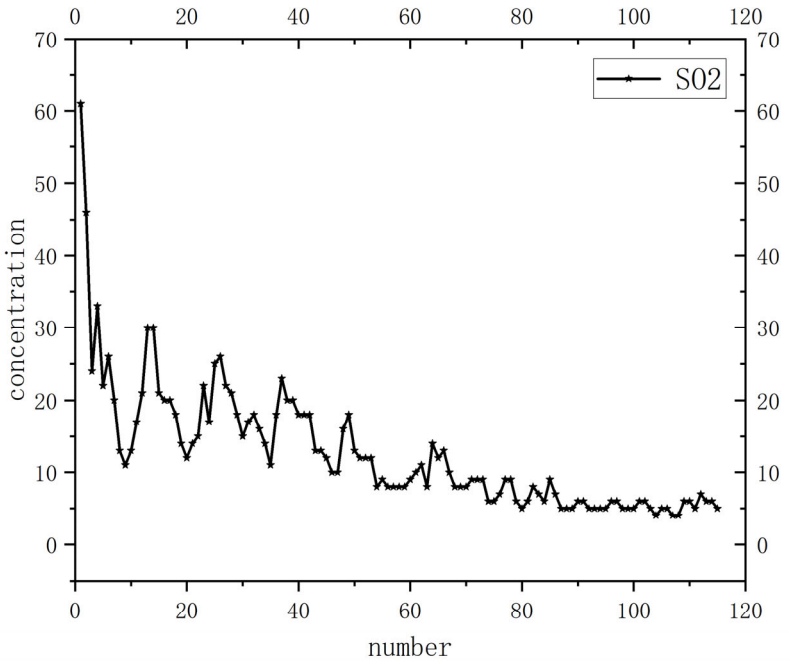


Figure 6: SO2 Concentration change chart

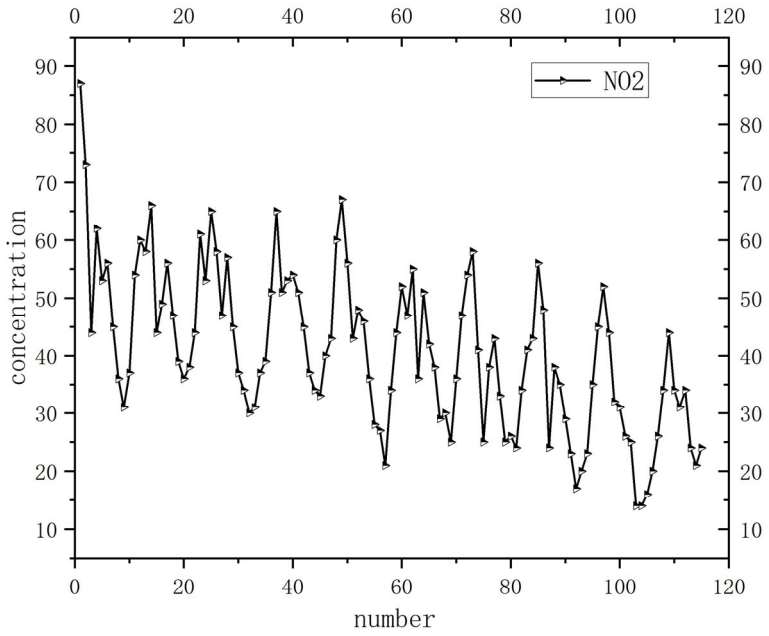


Figure 7: NO2 Concentration change chart

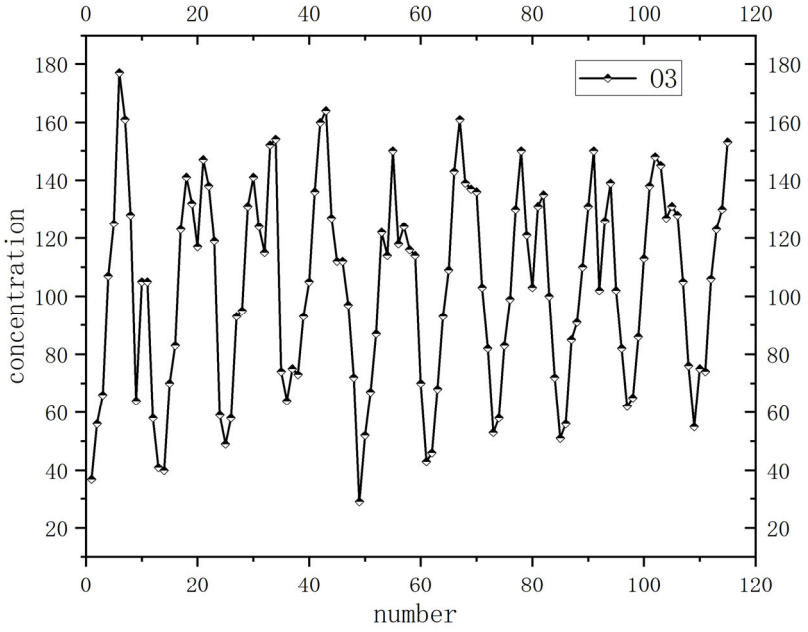


Figure 8: O3 Concentration change chart

Quantitatively analyse the significance of each pollutant indicator, therefore, this paper considers the establishment of linear regression fitting pollutant and AQI mathematical expressions, and the final fit goodness of fit $R^2$ as well as F and t-tests are analysed and illustrated.

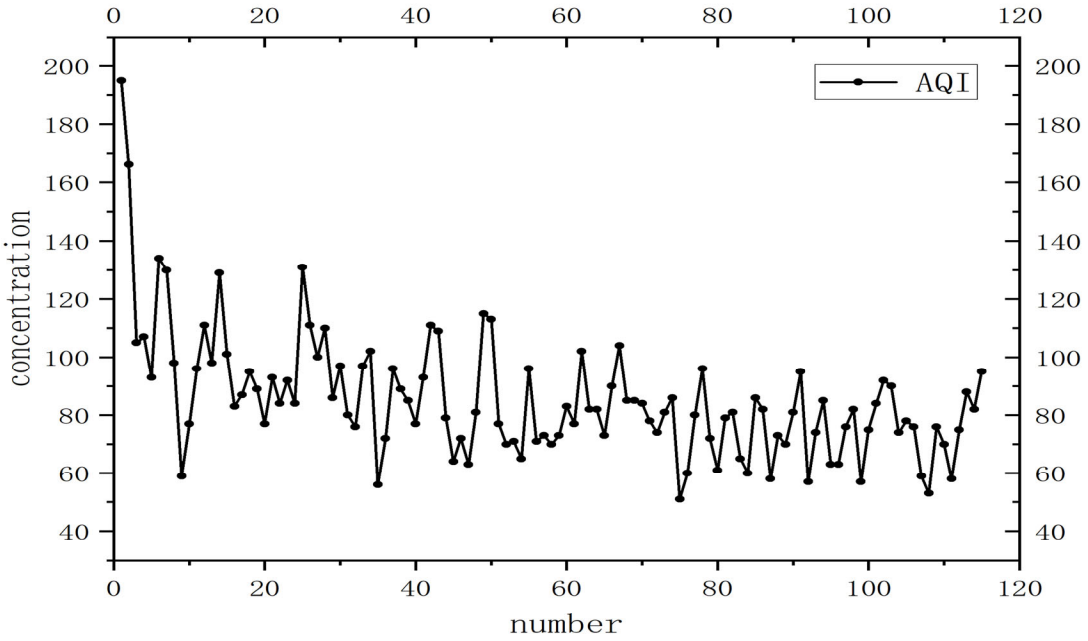


Figure 9 : AQI Concentration change chart

Let the linear regression model of random y with general variables $x_1, x_2, \cdots x_k$ be:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k + \varepsilon \tag{11}$$

Where $\beta_0, \beta_1, \cdots \beta_k$ is k+1 unknown parameters, $\beta_0$ is called the regression constant, $\beta_1, \cdots \beta_k$ is called the regression coefficient; y is called the explanatory variables; $x_1, x_2, \cdots x_k$ is k can be precise and can be controlled in general, called the explanatory variables, after the linear regression analysis, this paper comes up with the following table:

Table 1: Linear regression analysis table

| | Unstandardised coefficient | | Standardised coefficient | T inspect | P | $R^2$ | Adjust $R^2$ | F inspect |
|---|---|---|---|---|---|---|---|---|
| | B | Standard error | Beta | | | | | |
| C | 15.431 | 7.923 | None | 1.948 | 0.054 | | | |
| PM2.5 | 0.71 | 0.153 | 0.789 | 4.632 | 0.00 | | | |
| PM10 | 0.077 | 0.121 | 0.128 | 0.634 | 0.527 | | | |
| NO2 | -0.24 | 0.16 | -0.154 | -1.506 | 0.135 | 0.907 | 0.896 | F=76.152 |
| CO | 11.867 | 7.344 | 0.129 | 1.616 | 0.109 | | | |
| SO2 | 0.386 | 0.268 | 0.157 | 1.443 | 0.152 | | | |
| O3 | 0.273 | 0.038 | 0.445 | 7.228 | 0.00 | | | |

The analysis of the results of the F-test can be obtained that the significance P-value is close to 0, presenting significance at the level and rejecting the original hypothesis that its regression coefficient is 0. Therefore, the regression analysis model we constructed basically meets the requirements.

The formula of the model is as follows:

$$y = 15.431 + 0.71 * X1 + 0.077 * X2 - 0.24 * X3 + 11.867 * X4 + 0.386 * X5 + 0.273 * X6 \tag{12}$$

This paper is based on the construction of multiple linear regression of AQI and pollutants linear fitting, its fit superiority $R^2$>90%, proving that the model fitting effect is excellent, and secondly, through the F test as well as the t test, proving that the model construction is reasonable [10-13]. The constructed model is able to illustrate the influence of each variable on AQI with precision, and it is able to predict the future AQI indexes by multiple regression, and the establishment of this model has a certain degree of extensibility.

In summary, this paper further tests the multicollinearity between indicators in the multiple linear regression model by introducing the VIF variance inflation factor to test the multicollinearity, and if the test passes, it proves that there is no multicollinearity between the variables, i.e., the indicators are selected reasonably [14].

Variance inflation factor is a statistical indicator used to detect multicollinearity. It is used to assess the degree of covariance between each independent variable and the other

independent variables in a regression model. VIF is calculated by using a regression model for each independent variable as the dependent variable and the other independent variables as the independent variables in a regression analysis and calculating the $R^2$ (coefficient of determination) of the regression model [15]. VIF is then the reciprocal of $R^2$ (1 divided by $R^2$) and is used to indicate the degree of variance inflation of the independent variables. A larger value of VIF indicates stronger covariance between the independent variable and the other independent variables.

Therefore, VIF can be used to test the problem of multicollinearity in a regression model. Generally speaking, if the VIF value of an independent variable exceeds a set threshold (usually 5 or 10), it indicates that there is a strong covariance in that independent variable, which needs to be further dealt with. In this paper, we further solved for the variance inflation factor, VIF, between each variable in the following table:

Table 2: Table of variance inflation factors for VIF

| Type of pollutant | Variance inflation factor VIF |
|---|---|
| PM2.5 | 5.925 |
| PM10 | 6.692 |
| SO2 | 6.431 |
| O3 | 8.958 |
| CO | 9.714 |
| NO2 | 4.383 |

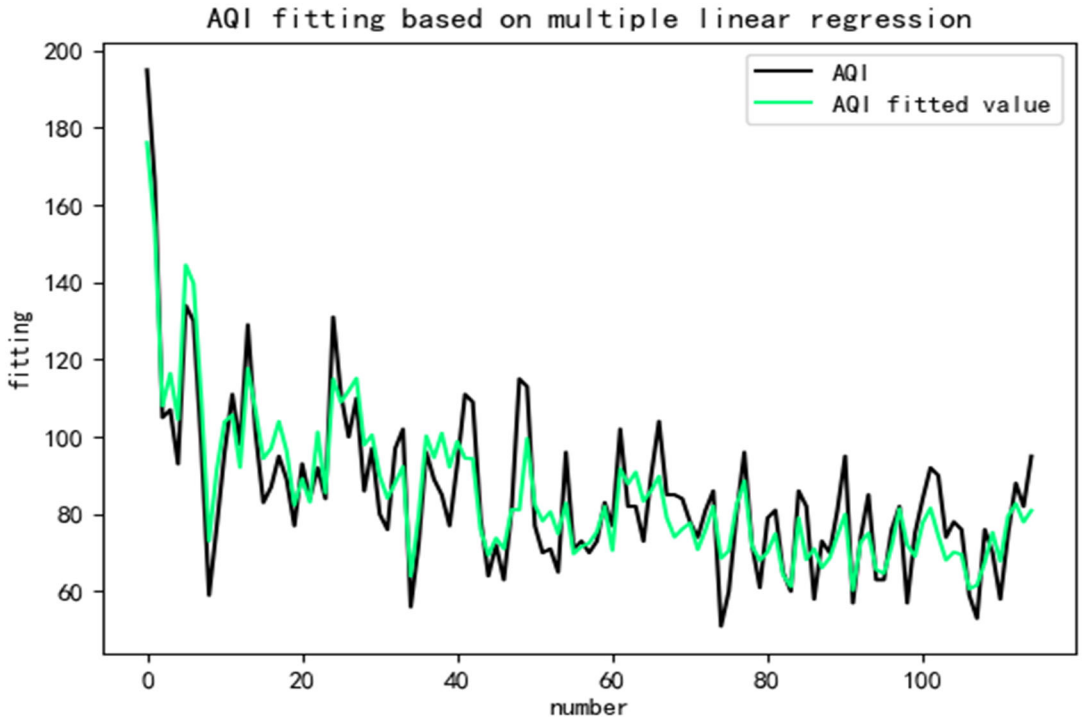


Figure 10: AQI fitting based on multiple linear regression

As can be seen from Table 2, there is no particularly strong multicollinearity between pollutant indicator types, and the variance expansion factor values are all in the range of 5-10, i.e., the model indicators selected in this paper are constructed reasonably.

### 3.1.4 Grey correlation analysis between pollutants and AQI parent series

Grey correlation analysis is a multi-factor statistical analysis method, the basic idea is to determine the degree of geometric similarity between the reference data column and a number of comparison data columns to determine whether they are closely linked, which reflects the degree of correlation between the curves

The steps of grey correlation analysis are as follows:

Determine the analysis sequence: choose a data series that can reflect the characteristics of the system's behaviour as the mother series (reference series), and choose the data series composed of factors affecting the system's behaviour as the sub-sequence (comparison series).

Data pre-processing: the original data are dimensionless, so that they are unified into an approximate range, commonly used methods include initialisation, averaging, intervalisation, and so on.

Calculate the grey correlation coefficient: Based on the absolute value of the difference between the parent series and the subsequence, calculate the correlation coefficient in each dimension, reflecting the degree of similarity between the two in that dimension. The commonly used formula is:

$$\xi_i(k) = \frac{\min\limits_{i,k} | x_0(k) - x_i(k) | + \rho \max\limits_{i,k} | x_0(k) - x_i(k) |}{| x_0(k) - x_i(k) | + \rho \max\limits_{i,k} | x_0(k) - x_i(k) |} \tag{13}$$

Where $x_0(k)$ is the value of the parent sequence in the kth dimension, $x_i(k)$ is the value of the ith subsequence in the kth dimension, and it is a resolution coefficient between 0 and 1, generally 0.5 [16].

Calculate the degree of association: for each subsequence, find the average value of the association coefficient in all dimensions as the degree of association between the subsequence and the parent sequence, reflecting the overall degree of similarity between the two [17]. The calculation formula is:

$$r_i = \frac{1}{n}\sum_{k=1}^{n} \xi_i(k) \tag{14}$$

Where n is the number of dimensions of each sequence.

In this paper, the following grey correlation coefficient table is derived by calculating the grey correlation degree of each column of alphabetical sequences and ranking them in terms of grey correlation:

Table 3: Grey correlation table

| Licence Holder | Relatedness | Rank |
|---|---|---|
| CO | 0.881 | 1 |
| NO2 | 0.853 | 2 |
| PM10 | 0.842 | 3 |
| O3 | 0.813 | 4 |
| PM2.5 | 0.811 | 5 |
| SO2 | 0.787 | 6 |

Combined with the correlation coefficient weighting process, the final correlation value table $x$, using the correlation value for the evaluation of the six evaluation objects for the evaluation of the ranking; correlation value between 0 ~ 1, the greater the value represents the stronger the correlation with the AQI indicators, that is, means that its evaluation is higher. The above table shows that for the six evaluation items, CO is the most highly rated (correlation: 0.881), followed by NO2 (correlation: 0.853).

In this paper, the pollutant index data of the relevant prefecture-level cities are obtained by checking the relevant websites of China Air Quality Index, and after the linear fitting of the multiple linear regression equations established on the index data to their air quality index AQI, the top 10 cities in the country are finally collated as the following table:

Table 4: Air Quality Rankings

| Cities | AQI |
|---|---|
| Dazhou | 16 |
| Zigong | 16 |
| Hegang | 17 |
| Heihe | 17 |
| Jiamusi | 17 |
| Shuangyashan | 17 |
| Yichun | 17 |
| Nanchong | 18 |
| Qiqihar | 18 |
| suining | 18 |

## 4 CONCLUSION

Based on the constructed regression equations, the air quality data collected from cities across the country were fitted to produce the ten cities with the best air quality: Dazhou City, Hegang City, Heihe City, Jiamusi City, Shuangyashan City, Yichun City, Nanchong City, Qiqihar City, and Suining City.

## REFERENCES

[1] WANG Huiwen,MENG Jie. A predictive modelling approach for multiple linear regression[J]. Journal of Beijing University of Aeronautics and Astronautics,2007 (04):500-504.

[2] Chunya Hu. Research and application of optimisation index system of aging mine ventilation system based on variance expansion factor [D]. China University of Mining and Technology,2016.

[3] Sun Yugang. Research on grey correlation analysis and its application[D]. Nanjing University of Aeronautics and Astronautics,2007.

[4] YANG Yongping,WU Dianfa,WANG Ningling. Comprehensive evaluation of thermal power unit performance based on combined weight-advantageous solution distance method[J]. Thermal Power Generation,2016,45(02):10-15.

[5] Pang Feng. Principle of simulated annealing algorithm and application of the algorithm to optimisation problems[D]. Jilin University,2006.

[6] Hong Lihua,Zhou Weihong,Huang Qionghui. Research on data visualisation based on Python[J]. Science and Technology Innovation and Application,2022,12(33):36-40.

[7] Zhang Hongxian,Ma Yaofeng. Multiple regression prediction of China's inbound tourism market[J]. Resource Development and Market, 2005, 21(2):2.

[8] Huang Geng-Wen,Yang Lian-Yue,Lu Wei-Qun,et al. Multiple regression analysis of complications after resection of hepatocellular carcinoma[J]. Chinese Journal of Practical Surgery, 2006, 26(6):2.

[9] XIN H, WANG Congxu, LU Wei, et al. Multiple regression analysis of the correlation law between Chinese medicine physical quality and quality of life in 663 cases of hypertension[J]. China Journal of Basic Chinese Medicine, 2011, 17(7):2.

[10] MA Feng, YANG Fajun, CHEN Runqiao, et al. Multiple regression analysis of the impact of groundwater mining on ground subsidence in Tianjin[J]. Chinese Journal of Geological Hazards and Prevention, 2008, 19(2):4.

[11] Zhang MY, Yan HF, Li CH, et al. A correlation study of neuropsychological and related examinations in Alzheimer's disease: multiple regression analysis of (4) cognitive function tests with BEAM, BEP, and CT [J]. Shanghai Psychiatry, 1996.

[12] LIU Zheng,DANG Guangzhe,LIU Xiaojun. Application of multiple regression analysis in the prediction of coal seam gas content[J]. Mining Safety and Environmental Protection, 2013(05):57-60.

[13] Zhou Liandi. Multiple regression analysis and its application in ship research and design[M]. National Defence Industry Press, 1979.

[14] FU Fengling, ZHOU Shufeng, PAN Guangtang, et al. Multiple regression analysis of drought tolerance coefficient of maize[J]. Journal of Crops, 2003, 29(3):5.

[15] Xie Yayu,Zhang Yukun. Correlation and multiple regression analyses of life event factors and mental health status of ethnic minority college students in China[J]. Chinese Journal of Mental Health, 1993, 007(004):182-184.

[16] Wei H. Application of multiple regression analysis in spatial prediction of landslides[J]. Chang'an University, 2011.

[17] Jaccard. Interaction in multiple regression[M]. Gezi Publishing House,2012.