Analysis of Influencing Factors of Air Quality in Shijiazhuang Based on Multiple Linear Regression Model

Faye Wang*
School of Business, Shandong Normal University, Jinan, Shandong, China
**Corresponding Author.*

Abstract: The main purpose of this report is to find out the main factors affecting the air quality in Shijiazhuang by analyzing the air quality data of Shijiazhuang in 6 years (2018-2023). In recent years, with the acceleration of industrialization and urbanization, air pollution has become increasingly serious in Shijiazhuang. By using multiple linear regression model, this study analyzed the effects of PM2.5, PM10, SO_2, CO, NO_2 and O_3_8h on AQI (air quality index), and revealed the contribution of each pollutant to air quality. The results of the regression model showed that PM2.5 was the most important factor affecting air quality, and its change was positively correlated with AQI. Secondly, the effects of PM10, CO and O_3_8h are also significant, which aggravate the air pollution in Shijiazhuang to a certain extent. Although NO_2 and SO_2 also have a certain impact on AQI, their impact is smaller than other factors. Through the analysis of these pollutants, this report provides data support and theoretical basis for urban air quality management and pollution source control.

Keywords: AQI; PM2.5; PM10; CO; O_3_8h

1. Preface

With the acceleration of industrialization and urbanization, the problem of air pollution is becoming more and more serious around the world, especially in developing countries and big cities. As the most populous country in the world, many large and medium-sized cities in China are facing severe air quality problems. With the increase of car ownership, coal combustion and industrial emissions, air pollutant emissions continue to rise, resulting in high concentrations of harmful particulate matter such as PM2.5 and PM10, which poses a great threat to the health and living environment of residents.

As the capital of Hebei Province, Shijiazhuang is located in the center of the North China Plain. The geographical location is special, and the air pollution problem is particularly serious.[1] Shijiazhuang is not only facing the pollution of industrial emissions and traffic exhaust, but also the pollution caused by the use of a large amount of coal during winter heating. According to the monitoring data for many years, the air quality in Shijiazhuang is usually at a poor level. Especially in winter, the peak concentration of PM2.5 often exceeds the national standard [2], which seriously affects the quality of life and health of residents.

In this context, the effective management and improvement of air quality urgently need scientific analysis and prediction methods. As a classical statistical analysis method, multivariate linear regression model has unique advantages in dealing with complex data relations, and can clearly reveal the internal relationship between various factors and air quality. By establishing a multiple linear regression model, we can quantify the influencing factors of air quality and analyze the mechanism of different pollutants (such as PM2.5, PM10, NO_2, SO_2, etc.) on air quality, so as to provide data support for policy makers. In this study, by collecting the air quality monitoring data of Shijiazhuang City, combined with relevant meteorological data, a multiple linear regression model was established to analyze the main influencing factors of air quality in Shijiazhuang, to explore the contribution of different factors to air pollution, and to predict the air quality in the next period of time. The research results can not only provide scientific basis for environmental protection decision-making in Shijiazhuang, but also provide reference for air quality prediction and improvement in other similar cities.

2. Research Methods

2.1 Data Source

This report is based on 1796 air quality data of Shijiazhuang from December 02,2013 to November 04,2018. In this report, the air quality index (AQI) [3] is selected as the dependent variable. By defining the scope of AQI, the air quality is divided into seven grades, which are excellent, good, mild pollution, moderate pollution, severe pollution, and severe pollution.

2.2 Indicator Design

After determining the dependent variables of this report, the explanatory variables are considered. In this report, PM2.5, PM10, SO2, CO, NO_2 and O_3_8h are selected as explanatory variables [4][5], which are explained as follows:

(1)AQI [6]: The value of comprehensive air quality is often used to describe the quality of air quality. The higher the value, the more serious the pollution. According to the AQI value range, the air quality is divided into six grades: excellent (0-50), good (51-100), mild pollution (101-150), moderate pollution (151-200), severe pollution (201-300), severe pollution (300+).

(2)PM2.5: Fine particulate matter, which indicates the concentration of particulate matter with a diameter of less than 2.5 microns in the air, has a significant impact on air quality and health.

(3)PM10: It is also a particulate matter, but has a large particle size, usually between 2.5 and 10 microns in diameter in the air.

(4)SO_2: The concentration of sulfur dioxide mainly comes from industrial activities and the combustion of fossil fuels. High concentration will affect air quality.

(5)CO: The concentration of carbon monoxide mainly comes from traffic and industrial emissions, which will cause harm to human health when the concentration is high.

(6)NO_2: The concentration of nitrogen dioxide, mainly from traffic, industrial emissions, etc., is toxic, and the higher the concentration, the greater the impact on air quality.

(7)O_3_8h: The 8-hour concentration of ozone has a dual effect on air quality. Low concentration is beneficial to prevent ultraviolet rays, but high concentration will form air pollution.

2.3 Analysis Method

This study uses multiple linear regression model [7] to explore the relationship between variables. The F test is used for the overall significance test to measure whether the regression model is generally statistically significant, and the t test is used to evaluate whether the impact of individual variables is significant. In addition, this study also calculated the adjusted R^2 to measure the goodness of fit of the model, and used the multicollinearity diagnostic (VIF) [8] test to test whether there is a strong collinearity between the explanatory variables to ensure the robustness of the regression analysis.

3. Research Results

3.1 Descriptive Statistics

Firstly, the specific descriptive statistical indicators are calculated for each variable, and the results are shown in Table 1.

Table 1. Descriptive Statistics of Each Variable

Variable Name	Minimum Value	Maximum Value	Mean Value	Standard Deviation	Variance
PM2.5 μg/㎥	0	621	92.29	78.359	6140.191
PM10 μg/㎥	0	866	159.88	108.809	11839.291
SO_2 μg/㎥	4	324	43.71	40.803	1664.923
CO mg/㎥	0	10	1.41	1.133	1.284
NO_2 μg/㎥	9	183	51.97	25.924	672.052
O_3_8hμg/㎥	0	297	93.27	59.407	3529.151
AQI μg/㎥	0	500	135.37	84.429	7128.332

In order to eliminate the influence of dimension, the above variables are standardized, as shown in Table 2. After standardization, the mean of all variables is 0, the standard deviation is 1, and the variance is 1. This allows variables of different units and dimensions to be compared directly.

Table 2. Standardized Descriptive Statistics of Each Variable

	Mean Value	Median	Standard Deviation	Variance	Minimum Value	Maximum Value
Zscore: PM2.5	0.00	-.31	1.00	1.00	-1.18	6.75
Zscore: PM10	0.00	-.25	1.00	1.00	-1.47	6.49
Zscore: SO_2	0.00	-.31	1.00	1.00	-.97	6.87
Zscore: CO	0.00	-.36	1.00	1.00	-1.16	7.93
Zscore: NO_2	0.00	-.19	1.00	1.00	-1.66	5.05

Zscore: O_3 8h	0.00	-.14	1.00	1.00	-1.57	3.43
Zscore: AQI	0.00	-.29	1.00	1.00	-1.60	4.32

Next, a simple histogram is made for each standardized explanatory variable to observe the shape of the data distribution, mainly to check whether there are obvious data anomalies. The results are shown in Figure 1-6.

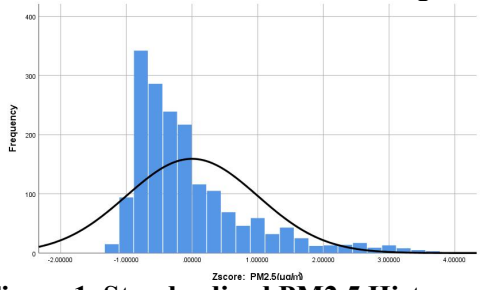

Figure 1. Standardized PM2.5 Histogram

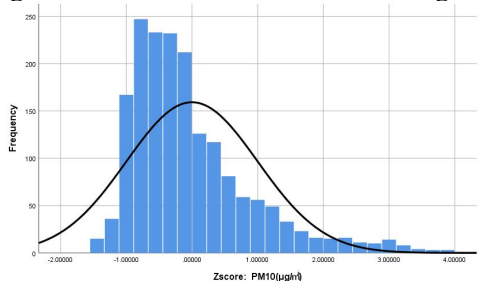

Figure 2. Standardized PM10 Histogram

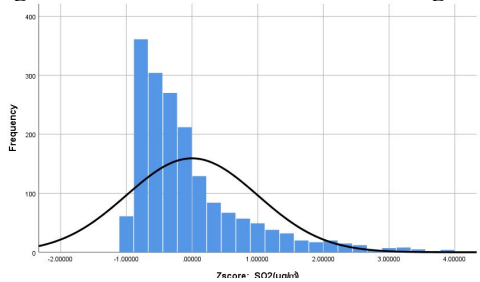

Figure 3. Standardized SO_2 Histogram

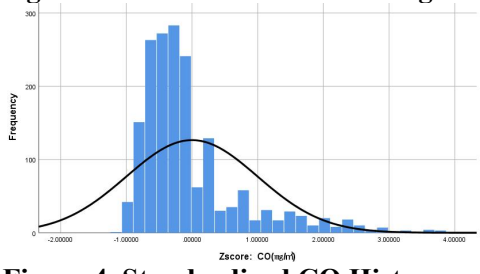

Figure 4. Standardized CO Histogram

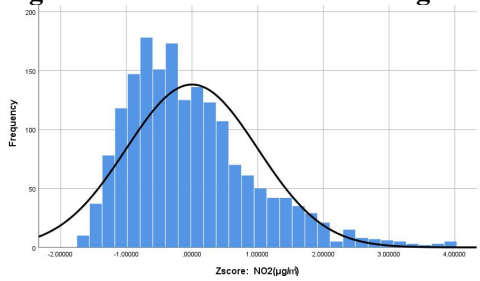

Figure 5. Standardized NO_2 Histogram

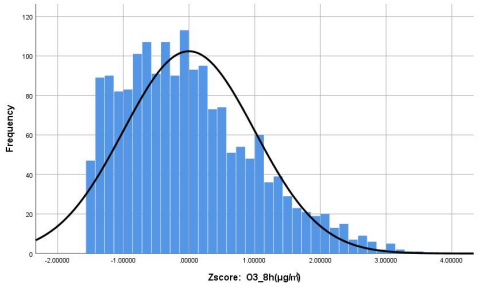

Figure 6. Standardized O_3_8h Histogram

Then, this report gives a brief description of the correlation between dependent variables and explanatory variables. Figure 7-12 shows the scatter relationship between the explanatory variables (PM2.5, PM10, SO_2, CO, NO_2, O_3 8h) and AQI.

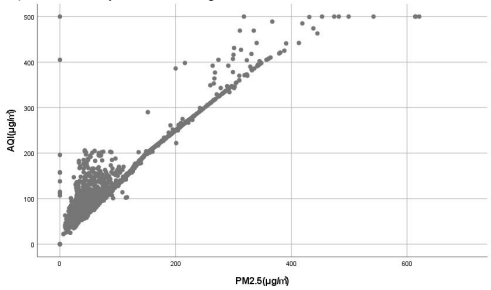

Figure 7. Scatter Plot of PM2.5 and AQI

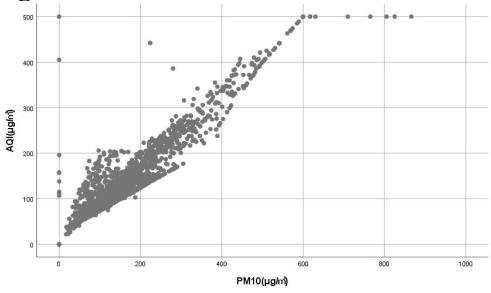

Figure 8. Scatter Plot of PM10 and AQI

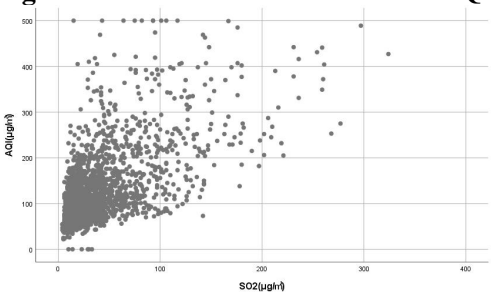

Figure 9. Scatter Plot of SO_2 and AQI

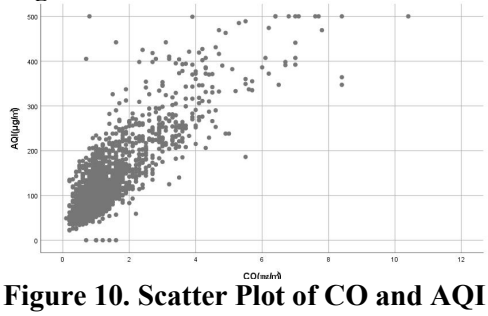

Figure 10. Scatter Plot of CO and AQI

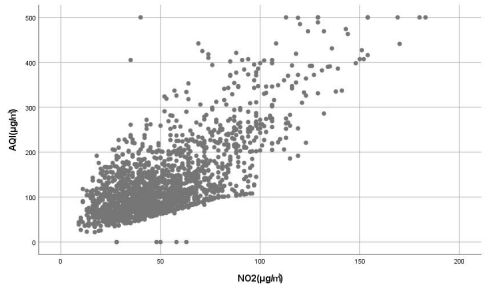

Figure 11. Scatter Plot of NO_2 and AQI

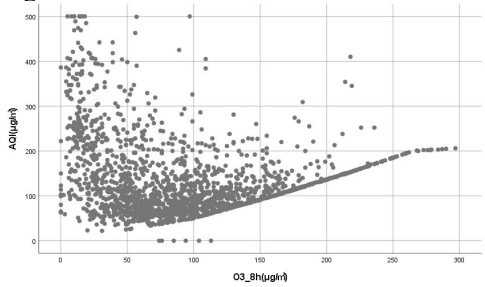

Figure 12. Scatter Plot of O_3_8h and AQI

Figure 13-18 shows the box plot between each standardized explanatory variable and the quality level, and further visually shows the distribution of variables and their outliers. The box plot can reveal the median, quartile and outliers of the data, and help us better understand the relationship between pollutant concentration and air quality.

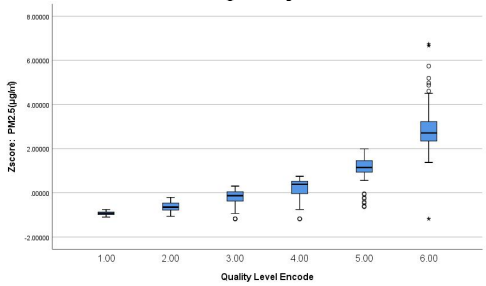

Figure 13. Boxplot of Standardized PM2.5 and AQI

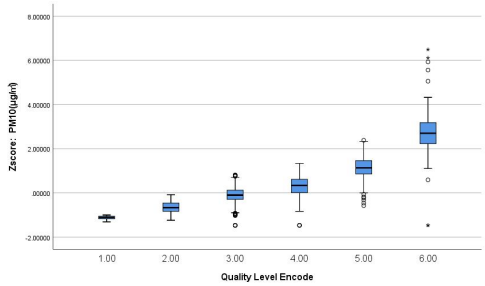

Figure 14. Boxplot of Standardized PM10 and AQI

3.2 Multiple Linear Regression

On the basis of descriptive analysis, this report will further analyze the relationship between AQI and each explanatory variable. Firstly, a simple multiple linear regression model is established for the data, and the relevant results are shown in Table 3, Table 4 and Table 5.

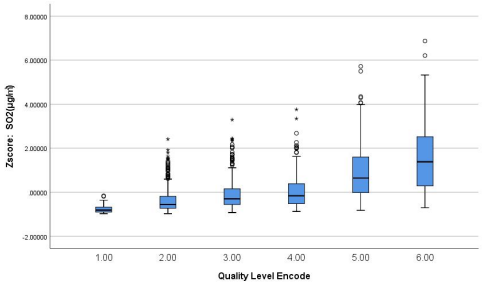

Figure 15. Boxplot of Standardized SO_2 and AQI

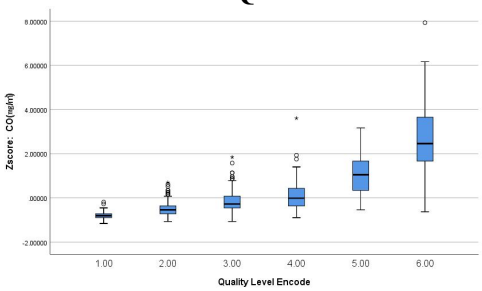

Figure 16. Boxplot of Standardized CO and AQI

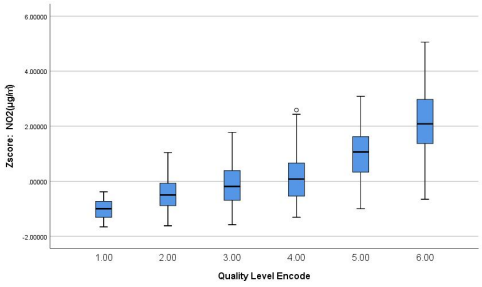

Figure 17. Boxplot of Standardized NO_2 and AQI

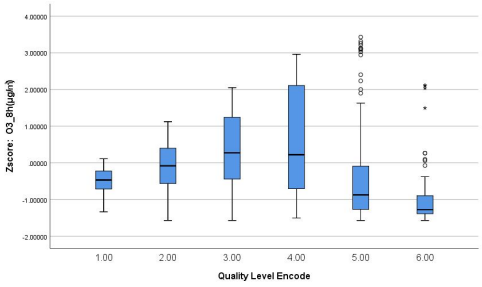

Figure 18. Boxplot of Standardized O_3_8h and AQI

Table 3. Model Summary

Model	R	R^2	Adjusted R^2	Errors in Standard Estimates	significance
1	.956	.914	.914	.294	.000

Table 4. ANOVA

Model	Squares	Degree of Freedom	Mean Square	F	Significant
Regression	1640.861	6	273.477	3174.073	.000
Residual	154.139	1789	.086		
Total	1795.000	1795			

Table 5. Regression Model Results

Model	Beta	t	Significant	VIF
Constant		.000	1.000	
Zscore: PM2.5	.670	24.425	.000	15.690
Zscore: PM10	.206	7.729	.000	14.799
Zscore: SO_2	.035	3.464	.001	2.150
Zscore: CO	.113	6.641	.000	6.083
Zscore: NO_2	.010	.719	.472	3.765
Zscore: O_3_8h	.164	20.496	.000	1.337

4. Analysis and Discussion

4.1 Descriptive Statistical Analysis

Firstly, Table 1 shows the descriptive statistics of air quality related variables in Shijiazhuang, including PM2.5, PM10, SO_2, CO, NO_2, O_3_8h and AQI. The following are the main statistical indicators of each variable:

(1)The minimum value of PM2.5 is 0, the maximum value is 621, the mean value is 92.29, and the standard deviation is 78.36, indicating that during the monitoring period, the PM2.5 concentration has large fluctuations in different time periods, and sometimes more serious pollution occurs. The variance is 6140.191, which further confirms the large volatility of the data.

(2)The mean value of PM10 was 159.88, and the standard deviation was 108.81. Although the minimum value of PM10 is also 0 and the maximum value is 866, the concentration of PM10 is higher than that of PM2.5 in general, indicating that Shijiazhuang City contains more large particles in the air.

(3)The mean value of SO_2 is 43.71, the standard deviation is 40.80, the maximum value is 324, and the minimum value is 4. This indicates that the concentration of SO_2 changes greatly compared with other pollutants, and the concentration of SO_2 in some areas may exceed the standard.

(4)The mean value of CO is 1.41, the standard deviation is 1.13, the maximum value is 10, and the minimum value is 0, indicating that the concentration of carbon monoxide is low, but there is still some volatility, which may be closely related to traffic emissions and winter heating.

(5)The mean value of NO_2 is 51.97, the standard deviation is 25.92, the maximum value is 183, the minimum value is 9, and the NO_2 concentration has a large range of variation compared with other variables.

(6)The mean value of O_3_8h was 93.27, the standard deviation was 59.41, and the maximum value was 297, indicating that the concentration of ozone fluctuated greatly in different time periods, especially in summer.

(7)The mean value of AQI is 135.37, the standard deviation is 84.43, the maximum value is 500, and the minimum value is 0. The fluctuation range of AQI value is large, and the extreme value appears in some periods, which shows the severity of air pollution.

Next, it can be seen from the histogram that the distribution of PM2.5, PM10, SO_2, CO, NO_2, and O_3_8h all show a right deviation, indicating that although the concentration of air pollutants is low in most periods, the concentration will increase significantly in a specific period of time, which may be related to specific pollution sources or weather conditions. Among them, the concentration of SO_2 and CO is higher, while the distribution of NO_2 and O_3_8h is the most dispersed, which may be related to the source and diffusion mechanism of pollutants.

Then, it is obvious from the scatter diagram that there is a positive correlation between the concentration of pollutants such as PM2.5, PM10, SO_2, NO_2 and AQI. When the concentration of these pollutants increases, the AQI value also tends to increase, indicating that these pollutants are an important cause of air quality deterioration. The relationship between PM2.5 and AQI is particularly significant. When the concentration of PM2.5 is high, the AQI value is usually higher; the relationship between NO_2 and CO and AQI is relatively weak, but a certain correlation can still be observed. The relationship between O_3_8h and AQI is more complex. Although AQI tends to be higher when ozone concentration is high, in some cases, high ozone concentration may be related to other factors (such as meteorological conditions).

Finally, the box line diagram was observed. The box line diagram of PM2.5 and PM10 showed that the changes of PM2.5 and PM10 concentrations were closely related to the fluctuation of AQI, and the abnormal values were obvious. The box plots of SO_2 and NO_2 show the fluctuations of these pollutants in extreme cases, suggesting that they may reach higher concentrations at some time points, resulting in a serious decline in air quality. The box plot of O_3_8h shows the seasonal fluctuation of ozone concentration, especially

in spring and summer, ozone concentration may lead to a significant increase in AQI.

4.2 Analysis of Multiple Linear Regression Results

In Table 3, R = 0.956 indicates that the model has a very high degree of fitting to AQI (air quality index), indicating that the predictive variables can well explain the changes in dependent variables. R square = 0.914 indicates that 91.4 % of the AQI variation can be explained by the selected six standardized explanatory variables (PM2.5, PM10, SO_2, CO, NO_2 and O_3_8h). The adjusted R square = 0.914 further verifies the robustness of the model, and there is not too much over-fitting due to the increase of variables. The F value is 3174.073, and the significance level p = 0.000, indicating that the regression model is significant as a whole, and the selected explanatory variables are statistically significant for the dependent variables.

In table 4, the sum of squared regression (SSR) is 1640.861, accounting for the vast majority of the total sum of squares, indicating that the model has strong explanatory power. The sum of squared residuals (SSE) is 154.139, which is relatively small, indicating that there is less variation that is not explained. The F value again verifies the overall significance of the model, and p = 0.000 indicates that the explanatory variables used are significantly correlated with AQI.

In Table 5, the unnormalized coefficient of the constant is close to 0, and the significance level is 1.000, indicating that the normalized data does not require constant offset. PM2.5 (β = 0.670) is the most influential factor on AQI, indicating that the concentration change of PM2.5 contributes the most to the air quality index, p = 0.000, which is significant. PM10 (β = 0.206) is a secondary important factor, PM10 concentration also has a significant positive impact on AQI, p = 0.000, significant; O_3 (β = 0.164) had a moderate effect on AQI, p = 0.000, and the significance was strong; CO (β = 0.113) had little effect on AQI, but it was still significant. SO_2 (β = 0.035) had the least impact on AQI, but it was still significant. The effect of NO_2 (β = 0.010) was very weak and not significant.

The maximum value of VIF is 15.690 (PM2.5), and the minimum value is 1.337 (O_3). The VIF values of most explanatory variables are within a reasonable range (less than 10). The VIF of PM2.5 and PM10 is higher, indicating that there may be strong collinearity between the two, but it is still within the acceptable range.

5. Conclusion

Through the multiple linear regression model, the corresponding regression coefficients obtained in this report can be used to predict the future AQI level. This report reveals the main factors affecting the air quality in Shijiazhuang. According to the regression model analysis, PM2.5 and PM10 have a greater impact on AQI. Therefore, reducing the emission of fine particulate matter (PM2.5) and inhalable particulate matter (PM10) should be the primary task of air pollution control. These pollutants can be reduced by strengthening industrial emission control, promoting clean energy, and reducing road traffic pollution. The regression coefficient of CO is large, indicating that it has a significant impact on AQI. Controlling CO emissions requires focusing on controlling traffic pollution, especially vehicle exhaust emissions; ozone concentration is often the main source of air pollution in summer. Therefore, special attention should be paid to ozone concentration in summer and corresponding control measures should be taken, such as reducing industrial emissions and limiting the use of certain emission sources.

This report provides a scientific basis for air pollution control in Shijiazhuang. Based on these analysis results, policy makers can focus on controlling the emission of pollutants such as PM2.5 and PM10, and take measures to reduce the concentration of CO and O_3_8h, so as to improve the air quality of the city and the living environment of residents.

References

[1] Wang Yixu. Analysis of PM2.5 pollution characteristics and driving differences in Beijing-Tianjin-Hebei region. Shandong Normal University, 2022.DOI: 10.27280 / d. cnki. gsdsu.2022.001636.

[2] Zhu Jiaming, Xi Haonan. Prediction and prevention of air pollution under the background of carbon peak. Journal of Yunnan University for Nationalities (Natural Science Edition), 1-7 [2025-02-09]. http: // kns.cnki.net/kcms/detail/53.1192.n.202412

31.1241.006.html.
[3] Wang Wenjuan. AQI analysis and prediction based on XGBoost-ARIMA-DT model. Anhui Normal University, 2024.DOI: 10.26920 / d. cnki. gansu. 2024.000195.
[4] Li Xuan, Liu Jia, Liu Xinyu. Prediction of air quality grade classification under multi-machine learning model. Desert and oasis meteorology, 2024,18 (06): 146-153.
[5] Wang Wenwen. Ambient air quality assessment and analysis of influencing factors. Leather production and environmental protection technology, 2024,5 (11): 164-166.DOI: 10.20025 / j. cnki.CN10-1679.2024-11-57.
[6] Zhang Shunshun, Lu Yanxi, Luo Wei, etc. AQI-based air pollutant prediction research. Energy and Environment, 2024, (01): 97-99 + 120.
[7] Liang Ying, Zhang Yuhai. Correct application and expression of multiple linear regression method. Chinese Journal of Child Health, 2020, 28 (02): 230-232.
[8] Liu Fang, Dong Fenyi. Research on the diagnosis and treatment of multicollinearity in econometrics. Journal of Zhongyuan University of Technology, 2020,31 (01): 44-48 + 55.

