Proceedings of the 3rd International Conference on Materials Chemistry and Environmental Engineering (CONF-MCEE 2023), Part II
DOI: 10.54254/2755-2721/7/20230354

Empirical analysis of air quality in Shanghai based on multiple linear regression model

Botao Cao

Economics Department, Shanghai Normal University, 100 Guilin Road, Shanghai, 200434, CHN

botaocao@mail.weber.edu

Abstract. In recent years, with the rapid development of social and economic development, scientific and technological progress, human production and living standards have achieved tremendous improvement, at the same time, a large number of exhaust gases and soot emissions have become important sources of pollution, industrial energy consumption and urban construction, etc. make China's air environment pollution problem increasingly serious. The air quality in China gradually presents regional pollution, especially in the Yangtze River Delta, Pearl River Delta, and Beijing-Tianjin-Hebei regions. This paper takes Shanghai as the research object and empirically analyzes the main influencing factors of its air quality, which can provide reference not only for managing the air quality of this city, but also for the air quality management of similar cities. This paper empirically investigates the influencing factors of air quality in Shanghai based on the Shanghai air quality monitoring data, including AQI (Air Quality Index) and PM2.5, PM10, NO_2, CO, SO_2 and O_3 pollutant concentrations, released by the Shanghai Bureau of Ecology and Environment from January 2015 to October 2021. This paper uses the knowledge of econometrics to establish a multiple linear regression model with the help of RStudio, which is continuously estimated, tested and improved on the basis of the primary model to analyze the degree of influence of pollutants on AQI, and also to analyze the main pollutants affecting air quality by testing. The statistical results show that the primary air pollutants in Shanghai are O_3 and NO_2. Based on the analysis results, this paper puts forward corresponding suggestions about improving the air quality and protecting the ecological environment in Shanghai.

Keywords: AQI, Air Pollution, multiple linear regression model, Shanghai.

1. Introduction

In recent years, due to the continuous changes in the economy and society, the level of science and technology has leaped, the urbanization process has accelerated, and urban construction with high energy consumption and high emissions has made the problem of air pollution increasingly serious. The impact of human production and life on air quality cannot be ignored. Air quality has become one of the research topics of widespread concern in China. For example, some scholars believe that the quality of urban air quality is not only related to urban topography and urban construction but also seasonal and meteorological conditions are the primary influencing factors. This type of literature uses the monitoring data of air quality index and meteorological elements such as wind direction, wind speed, rainfall, air pressure, temperature, and cloud cover to explore the direct impact of meteorological conditions on air quality. Such studies have shown that in the absence of significant

DOI: 10.54254/2755-2721/7/20230354

changes in pollution source emissions, meteorological conditions directly affect air quality, and meteorological factors such as temperature, low cloud cover, average wind speed, and precipitation have a significant negative correlation with air quality. Combined with conventional weather forecast products and commonly used numerical forecast products, the multiple regression prediction equations of air quality meteorological factors are given, and air quality forecasts are made in combination with meteorological forecast products. The advantage of changing the law is that a clear relationship between natural factors can be obtained, but the pollution of human factors to the atmospheric environment is not involved. Environmental protection bureaus across China publish their relevant values in real-time, which are important indicators of environmental monitoring. The daily monitoring and evaluation of air quality based on the AQI index, and the quantitative description of air quality, have guiding significance for production and life. The AQI index provides a decision-making basis for the formulation of environmental policies and pollution control measures, to achieve the ultimate goal of improving the environment and promoting people's health.

In this paper, the concentrations of PM2.5, PM10, NO_2, CO, SO_2, O_3, and other pollutants are included in the empirical model, and six pollutant concentration variables are used to analyze their impact on AQI. Estimate, test and improve to verify the impact of pollutants on the air quality index, and analyze the main pollutants that affect air quality through testing. In this way, in addition to natural factors, the role of human factors in affecting the quality of the atmospheric environment can be clearly and intuitively studied.

2. Literature Review

The issue of air quality has always been concerned by countries around the world. With the development of China’s industrialization and urbanization, the consumption of energy and resources has been continuing to increase, causing serious air pollution problems. As the central city of the Yangtze River Delta, Shanghai has always been highly concerned about its air quality. In recent years, with the industrial upgrading of Shanghai, the implementation of relevant sustainable development policies, and the improvement of public awareness of environmental protection, the air quality in Shanghai and its surrounding areas have been significantly improved. But the control of air pollution is a long-term and systematic process. At present, there are few papers about the air quality in Shanghai. Therefore, when reviewing the literature, I also refer to the air pollution control reports of other cities. In general, this literature mainly focuses on two aspects: the analysis of factors affecting air quality and the prediction of future air quality. The relevant researches are mainly based on the daily data of 56 air quality stations in the city released by the Shanghai environmental protection department. Based on the air quality data during the Spring Festival in Beijing from 2016 to 2020 [1]. The authors used a multiple linear regression model to conclude that the influence of meteorological factors on the concentration of pollutants is dominant, and the level of social and economic activities also has a significant impact on air quality. Based on the daily data of Beijing from 2014 to 2017 [2]. They concluded that meteorological conditions, PM2.5 and O_3 have an important influence on the generation of haze weather. There was a study also found that economic development and public transportation development are mostly proportional to urban air quality, and the development of urban greening has no significant impact on urban air quality [3]. Domestic research in China on air quality prediction based on statistical prediction models and machine learning methods is relatively abundant. For example, the research pointed out that the multiple regression model based on meteorological factors and pollution concentration has a high accuracy for PM2.5 concentration prediction [4]; The research realized the prediction of PM2.5 indicators in Wuhan through the stepwise regression model and achieved good results [5,6]. summarized the application progress of deep learning methods in air quality forecasting, and pointed out that the existing machine learning methods can realize the prediction of air quality. Effective prediction, but its prediction accuracy can still be greatly improved, and gave a prospect for building a new deep learning model; Learning classification methods for empirical research, it is concluded that the prediction results of the decision tree model are inferior to the random forest model but better than the discriminant analysis and support vector machine models,

DOI: 10.54254/2755-2721/7/20230354

and the classification results can be displayed more clearly [7]. In general, machine learning models have higher prediction accuracy than general statistical models, but there are problems such as more complex model theory, more difficult model implementation, and poor interpretability. Therefore, this paper establishes a multiple linear regression model with the help of R, analyzes the impact of different pollutants on Shanghai air quality, and puts forward relevant policy suggestions.

3. Research Methodology

3.1. Data Sources

The data included in the explanatory variables selected in this paper are from the National Bureau of Statistics: *China Statistical Yearbook*, and the air quality and meteorological related data of Shanghai from January 2015 to October 2021 are selected in this paper. The data type is time series data, and the data are obtained from the Shanghai Bureau of Ecology and Environment, the National Oceanic and Atmospheric Administration (NOAA), and the National Center for Environmental Information (NCEI), with a total sample size of 2490.

3.2. Description of data the processing

Among the air quality index data, AQI reflects the comprehensive air quality, so this index is used as the target variable. According to the daily average AQI values, the air quality can be classified into 6 levels: excellent, good, lightly polluted, moderately polluted, heavily polluted, and severely polluted. The AQI ranges for these six levels are 0-50, 50-100, 101-150, 151-200, 201-300, and over 300, respectively. Since the data selected in this paper are natural weather data, which have certain randomness and uncontrollability, EXCEL software is used to pre-process the sample data and eliminate individual extreme sample data, so that the sample data as a whole are more stable and the regression and research results are more reliable.

3.3. Model selection and establishment

Taking the daily air quality data and meteorological data of Shanghai from January 2015 to October 2021 as samples, the sample data is analyzed by R, a preliminary model was established, and the least squares method was used for regression. According to the sample data, the estimator of the parameters of the simple linear regression model can be obtained by using the least squares estimator. Taking PM2.5, PM10, NO_2, CO, SO_2 and O_3 as the six independent variables and AQI as the dependent variable, a multiple linear regression equation was established, the data were sorted, and the air quality index was quantitatively analyzed. According to the theoretical analysis results of the correlation between air quality index and pollutants in the air, the six influencing factors and AQI index can be approximately assimilated into a linear correlation. According to the conclusion, the multiple linear regression equation model is established as follows:

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i \quad (1)$$

Where Y is the explanatory variable (air quality index); β_0 is a constant term with no particular significance; β_1, β_2, β_3, β_4, β_5 and β_6 are the parameters to be estimated; X_{1i} is the concentration of PM2.5 on day i, X_{2i} is the concentration of PM10 on day i, X_{3i} is the concentration of NO_2 on day i, X_{4i} is the concentration of CO on day i, X_{5i} is the concentration of SO_2 on day i, X_{6i} is the concentration of O_3 on day i, ε_i is the random error term.

Since the more serious the environmental pollution, the higher the AQI value, and PM2.5, PM10, NO_2, CO, SO_2 and O_3 are all pollutants, the higher the concentration of pollutants, the worse the environment, so the preset symbols of β_1, β_2, β_3, β_4, β_5 and β_6 are all positive.

$$Y = \beta_0 + \overset{+}{\beta_1} X_{1i} + \overset{+}{\beta_2} X_{2i} + \overset{+}{\beta_3} X_{3i} + \overset{+}{\beta_4} X_{4i} + \overset{+}{\beta_5} X_{5i} + \overset{+}{\beta_6} X_{6i} + \varepsilon_i \quad (2)$$

DOI: 10.54254/2755-2721/7/20230354

3.4. Descriptive Statistics for Variables

Table 1. Descriptive statistics for variables, unit: mg/m³.

Var	Min	Max	Mean	Median	S.D.	Dispersion coefficient
Y	21.00	206.00	64.80	60.00	26.50	0.41
X1	5.00	156.00	29.65	24.50	19.26	0.65
X2	11.00	311.00	45.11	37.00	30.18	0.67
X3	4.00	97.00	35.55	33.00	16.02	0.45
X4	0.30	1.50	0.61	0.60	0.18	0.30
X5	4.00	21.00	5.78	5.00	2.01	0.35
X6	14.00	205.00	89.17	82.00	35.91	0.40

According to the mean and median of the seven groups of data described in Table 1, the means of *Y, X_1, X_2, X_3, X_4, X_5,* and X_6 are greater than the medians, indicating that the seven groups of data are generally large. The representativeness or stability of the overall mean index (generally the mean) of the sample data of different variables can be illustrated by the comparison of the size of the dispersion coefficient. It can be seen from Table 1 that the dispersion coefficients of *Y, X_1, X_2, X_3, X_4, X_5,* and X_6 are all small, indicating that the 7 groups of data have good representation or strong stability; at the same time, their dispersion coefficients are all less than 0.7, indicating that the data distribution is relatively close.

4. Empirical Results Analysis

4.1. OLS estimation results

The least squares method (OLS) is used to estimate the model parameters, and the mathematical optimization technique is used to obtain the minimized sum of squared errors to find the best matching function expression by using RStudio. The estimated model parameter values are shown in Table 2. From the OLS regression results, the estimated model can be obtained as:

$$\widehat{AQI} = 0.4305PM2.5 + 0.2164PM10 + 0.1436SO_2 + 0.3445NO_2 + 0.4063O_3 + 13.1527CO - 5.875622 \quad (3)$$

$$R^2 = 0.8438 \; \bar{R}^2 = 0.8434 \quad (4)$$

According to the regression results:

The goodness of fit $R^2 = 0.8438$ is high, indicating that the regression estimation equation has a good fit. For each unit increase in the concentration of PM2.5, the air quality index will increase by 0.4305, and the significance is strong; for each increase in the concentration of PM10 by one unit, the air quality index will increase by 0.2164, and the significance is strong; the concentration of SO_2 increases by one unit, The air quality index will increase by 0.1436, and it is significantly stronger; the concentration of NO_2 increases by one unit, the air quality index will increase by 0.3445, and the significance is strong; the concentration of O_3 increases by one unit, the air quality index will increase by 0.4063, And the significance is strong; for each unit increase of CO concentration, the air quality index will increase by 13.1527, and the significance is strong; all meet the above parameter assumptions, which also shows that the main 6 air pollutants are all affecting the air quality in Shanghai.

DOI: 10.54254/2755-2721/7/20230354

Table 2. OLS preliminary regression results.

Call:
lm(formula = AQI ~ PM2_5 + PM10 + SO_2 + NO_2 + O_3 + CO)
Residuals:

Min	IQ	3Q	Max	Median
-40.32	-7.071	6.558	58.656	-0.77

Coefficients:	Estimate	Std. Error	t value	Pr(>| t |)	
(Intercept)	-5.875622	0.502115	-11.702	< 2e-16	***
PM2_5	0.430474	0.027019	15.932	< 2e-16	***
PM10	0.216422	0.016011	13.517	< 2e-16	***
SO_2	0.14359	0.083945	1.711	0.0873	.
NO_2	0.344486	0.023694	14.539	< 2e-16	***
O_3	0.406317	0.007782	52.21	< 2e-16	***
CO	13.152743	2.462898	5.34	1.01e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.1 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 11.4 on 2482 degrees of freedom
Multiple R-squared: 0.8438, Adjusted R- aquared: 0.8434
F-statistic: 2235 on 6 and 2482 DF, p-value: < 2.2e-16

4.2. Lagrange Multipliers: Serial Correlation Tests

Because the air quality data cannot determine the number of residual lag periods and the sample size is large, multiple LM tests are performed to determine the number of lag periods. Taking p = 0.05 as the standard, it is known from Table 3 that when the number of lag periods is 13, the p-value is less than 0.05. Therefore, the optimal number of lag periods is 13 periods.

Firstly, establish the auxiliary regression equation:

$$e_t = \gamma_1 e_{t-1} + \gamma_2 e_{t-2} + \cdots + \gamma_7 e_{t-7} + (\delta_0 + \delta_1 PM2.5 + \delta_2 PM10 + \delta_3 SO_2 + \delta_4 NO_2 + \delta_5 O_3 + \delta_6 CO) + u_t \quad (5)$$

$$\alpha = 0.05\ NR^2 = 600.1839\ df = 13 \quad (6)$$

Secondly, establish the hypothesis:

$$H_0: \gamma_1 = \gamma_2 = \cdots = \gamma_{13} = 0 \quad (7)$$

$$H_A: H_0\ is\ not\ true \quad (8)$$

Check the table to find out:

The critical chi-square value is 14.07, $NR^2 > 14.07$ so reject H_0..

Therefore, there is a serious thirteenth-period serial correlation in this equation.

Table 3. LM lag period test.

Breusch-Godfrey test for serial correlation of order up to 12
data: fit1
LM test = 20.818, df = 12, p-value = 0.05311
Breusch-Godfrey test for serial correlation of order up to 13
data: fit1
LM test = 23.848, df = 13, p-value = 0.03256

DOI: 10.54254/2755-2721/7/20230354

Since time series data are generated by random processes, and as far as air pollutants are concerned, the air pollution situation of the previous day will inevitably have an impact on the air pollution situation of the following day. Therefore, it has been tested that the air quality of a certain day may affect the air pollution situation. and air pollution in the following 13 days.

4.3. Significance Analysis

4.3.1. t-Test. From Figure 1, it is very clear that t1=15.932, t2=13.517, t3=1.711, t4=14.539, t5=52.210, and t6=5.340. A t-test was performed on each variable in the regression equation. The p-values of the explanatory variables PM2.5, PM10, NO_2, CO, and O_3 are extremely small, so the null hypothesis is rejected, and because β1; β2; β3; β4; CO, and O_3 had significant effects on AQI, while SO2's p-value was 0.0873, which was significant at the 10% significance level, but had less effect on AQI.

4.3.2. F-Test. To test the overall significance of the equation, under the assumption of a one-sided test with a given significance level of 5%, all explanatory variables PM2.5, PM10, NO_2, CO, SO_2, and O_3 are selected for the joint F test, and the results are shown in Table 4. Because F-statistic=2234.7, p-value = 2.2e - 16, so H_0 is rejected. The overall regression equation is significant, that is, there is a significant linear relationship between PM2.5, PM10, NO_2, CO, SO_2, O_3, and AQI.

Table 4. F-Test—Equation Joint Significance Test.

Linear hypothesis test						
Hypothesis:						
PM2_5 = 0						
PM10 = 0						
SO_2 = 0						
NO_2 = 0						
O_3 = 0						
CO = 0						
Model 1: restricted model						
Model 2: AQI - PM2_5 + PM10 + SO_2 + NO_2 + O_3 + CO						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2488	2064783				
2	2482	322514	6	1742269	2234.7	2.20E-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

5. Research Conclusion and Policy Suggestions

5.1. Research Conclusion

After the above multiple linear regression as well as hypothesis testing, it can be concluded that the main factor affecting the air quality in Shanghai is the concentration of the atmospheric pollutants NO_2, SO_2, and O_3, so air treatment is carried out mainly to control the concentration of NO_2, SO_2, and O_3 in Shanghai.

5.2. Suggestions Related to Improving the Air Quality in Shanghai

5.2.1. Reform the energy structure, promote industrial restructuring and develop new energy sources and clean, renewable resources. NO_2 is the key factor affecting air quality in Shanghai. Therefore, the Shanghai government and enterprises can increase the proportion of clean energy use in the energy structure, reduce the use of fossil fuels with high sulfur content such as coal, and replace coal burning

DOI: 10.54254/2755-2721/7/20230354

with gas. At the same time, the development of new energy sources advocates the use of wind, solar and other renewable green energy. Promote supply-side structural reform and accelerate industrial restructuring, pay attention to the development of low-pollution industries, expand technological capital investment, and improve combustion and recycling treatment technologies.

5.2.2. Control emissions and make full use of the self-cleaning capacity of the atmosphere, rationalize urban layout, and strengthen regional pollution investigation and remediation. Meteorological conditions affect the degree of atmospheric capacity for pollutants. Therefore, for windy, well-ventilated, convective areas and time periods that can accept more factory and mining enterprise activities, the atmosphere has a stronger ability to diffuse and dilute. Reasonable planning and arrangement of the urban and industrial area layout to avoid excessive concentration. Strengthen the standardized management of pollution emissions, to ensure that enterprises with coal meet the standards, the key areas of scientific and precise control, careful study of the sources and characteristics of pollutants, and strictly by the law to implement the assessment and accountability.

5.2.3. Increase green protection publicity, raise public awareness of the environment, advocate traffic restrictions, ban on burning and dust suppression, and actively improve air quality in the long term. Human factors are the main cause of ambient air pollution. The government should strengthen the supervision and publicity of environmental protection, raise people's awareness of protection, and fundamentally solve the problem of environmental management. Implement the new development concept, advocate green production and life, and take public transportation. It should advocate traffic restrictions, ban burning and discharge, strictly prohibit open burning of straw and garbage, improve the networked supervision system, and operate an efficient treatment mechanism to effectively reduce pollutants. And increase urban road cleaning efforts, for the implementation of bare ground greening or hardening, covering, and other dust suppression measures. In the long run, use the above measures to improve air quality in Shanghai with high efficiency and quality.

References

[1] YAO Yi, et al. Analysis on the influencing factors of air quality during the Spring Festival in Beijing in 2020. *Environment and Sustainable Development*, 46.02(2021):107-114. doi:10.19758/j.cnki.issn1673-288x.202102107.

[2] XU Changri, et al. Analysis of the influencing factors of haze in Beijing based on pollutants and meteorological elements. *Electric Power Environmental Protection* 37.01(2021):1-8.

[3] WANG Binhui and Wang Shu. Empirical study on factors affecting China's urban air quality——Evidence from the panel data of China's 31 major cities. *Journal of Fujian Agriculture and Forestry University (Philosophy and Social Sciences)*, 18.06(2015):29-33. doi:10.13322/j.cnki.fjsk.2015.06.007.

[4] WANG Juan. The Influencing Factors of Air Quality in Taiyuan. *Shanxi Science Technology*, 35.01(2020):58-61+64.

[5] LIU Huijun. Developing Pattern prediction, casual analysis and simulation of PM2.5 in Wuhan City. 2014. Hunan University, MA thesis.

[6] ZHU Yanmin,et al.. New Progress for Air Quality Forecasting Methods Based on Deep Learning. *Environmental Monitoring in China*, 36.03(2020):10-18.doi:10.19316/j.issn.1002-6002.2020.03.02.

[7] ZHAO Meng. Research on Atmospheric Environment Prediction Based on Data Mining Technology. 2017. Beijing Jiaotong University, MA thesis.