Analysis of Factors Influencing Air Quality in Chinese Cities Based on Multiple Linear Regression

Yijie Wang

{2023302011112@whu.edu.cn}

Mathematics and Statistics College, Wuhan University, Wuhan, 430072, China

Abstract. Based on the public panel data from several large cities in China, this paper investigates the effects of social and economic factors as well as weather situations in those cities. Through establishing a multiple linear regression model, taking into account the multi-collinearity and heteroscedasticity in the data, it was determined that industrial activities, energy utilisation, and increased wind speed all have varying degrees of positive correlations with PM2.5 concentrations; however, higher wind speeds will lead to some degree of decline in these concentrations. Therefore, it is determined that these components did not experience the same degree of impact across all areas. Based on the above research, new empirical data have been obtained to explore the causes of urban air pollution problems, and it is suggested that in order to drive sustainable development of cities, regionally adaptive strategies for air quality management should be developed.

Keywords: PM2.5; Air quality; Multiple linear regression; Chinese cities; Environmental policy

1 Introduction

Among various pollutants in China, fine particulate matter (PM2.5) is considered the most harmful because it can penetrate deep into the human respiratory system [1]. A long time of being in an environment with a large amount of PM2.5 can significantly increase the probability of having cardiovascular and cerebrovascular disease, as well as other respiratory illnesses, and ultimately die due to these ailments.

In China, with rapid economic development, industrialisation and urbanisation have all led to serious air pollution problems. The traditional production model of the industry relies heavily on fossil fuels; The number of cars has increased rapidly, leading to frequent haze weather in many cities around the world [2]. Empirical studies show that, in terms of driving factors of PM2.5 concentrations across cities in China; industrial processes and fossil fuel burning constitute primary sources [3].

To this end, the Chinese government has introduced a series of strict environmental protection policies since the end of 2010, such as the "Three-Year Action Plan for Winning the Blue Sky Defense Battle" released in 2018 [4]. Although some of these means have brought about a certain improvement in air quality at various levels across more parts of the country or temporarily reduced concentrations of PM2.5, many other cities remain seriously polluted.

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365116

Therefore, comprehensively and quantitatively understanding the determinants of urban air quality is essential for formulating effective and region-specific pollution control policies [5]. Existing studies have made in-depth explorations into the determinants of air quality in China. Many empirical studies show that, among other things, the industrial structure, degree of reliance on fossil energy use for production and daily life, as well as a large-scale urban economy; determine high levels of PM2.5 emissions [3,6]. Comparative studies of air quality index systems further highlight that social and economic factors explain the space differences in urban pollution patterns [7].

In recent years, research on the formation process of air pollution has shown that it is an extremely complicated nonlinear interaction among multiple socio-economic and meteorological factors [8]. Based on the above conclusions, it cannot be explained solely with traditional single-factor theory; Therefore, the causes of changes in city air pollution need to be explored from multiple dimensions.

From a methodological perspective, multiple linear regression is still widely used in research on air pollution to this day because it is clear and understandable. Methodologically, regression analysis requires a multicollinearity test and a heteroscedasticity test to be conducted first, otherwise it is difficult to draw dependable conclusions [9]; Otherwise, valid results cannot be obtained. In recent years, empirical research based on panel-data regression methods at home and abroad has explored the effects of human-emitted pollutants and meteorological elements on variations in PM2.5 concentration [10].

Although these achievements have been reflected upon, they are still somewhat limited. Some research based on old datasets; Some do not consider regional differences or the robustness of the method. The above defects have prompted us to conduct an updated empirical study based on the most recent data and strict regression analysis methods.

In view of the above problems, this study first constructs a panel dataset of key cities in China by integrating recently available PM2.5 concentration data and related socio-economic and meteorological variables. Subsequently, a multivariate linear regression model was established and rigorously validated to determine the size of the impact of key factors. Finally, based on the empirical research results, policies were proposed to provide a reference for air pollution control in various parts of China.

2 Methodology

2.1 Data Sources

This research combines several datasets available on Kaggle into an intact panel dataset. The source of the PM2.5 concentration mainly comes from two sets of datasets from Kaggle: "PM2.5 Data of Five Chinese Cities" and "Air Pollution in China (2015-2025)", which provide daily and annual PM2.5 concentrations for five major cities in China - Beijing, Shanghai, Guangzhou, Chengdu, and Shenyang. This set of data can provide a good reference for both cross-sectional and time-series analyses of urban air pollution.

The same set of data is collected to obtain the values of wind speed, temperature and relative humidity, so that all meteorological variables have the same sampling frequency both in time and space. Integrating socio-economic indicators, including GDP, the share of industrial output, and energy consumption per unit of output, at this time, city-level statistical data or data that have been matched with air quality datasets are used.

Using R, all datasets are cleaned and aggregated into an annual city-level format for easy panel regression analyses.

2.2 Variable Definition

This study uses annual average PM2.5 concentrations (in micrograms per cubic meter) as a dependent variable for each city. The independent variables are GDP per capita (an indicator of economic development), a share of industrial output (an indicator of the level of industrial activity), energy consumption intensity (an indicator of dependence on fossil fuels), and wind speed (an indicator of meteorological dispersion conditions). These variables have been widely used in empirical studies of air pollution and are described in theory about factors influencing the formation and diffusion of PM2.5.

2.3 Model Specification

To quantify the relationship between PM2.5 concentrations and influencing factors, the following form of a multiple linear regression model will be specified:

$$PM2.5_it = \beta 0 + \beta 1GDP_it + \beta 2Industry_it + \beta 3Energy_it + \beta 4Wind_it + \varepsilon_it \tag{1}$$

where i denotes the city, t denotes the year, and $\boldsymbol{\varepsilon_i t}$ is the error term. This modeling framework has been widely applied in air pollution studies to isolate socio-economic and meteorological effects [9,10].

2.4 Estimation and Diagnostic Tests

The regression analysis is conducted using R. Variance inflation factor (VIF) tests are applied to detect multicollinearity, while heteroskedasticity is addressed using robust standard errors where necessary. This diagnostic strategy is consistent with best practices in applied regression analysis [9].

3 Results

3.1 Descriptive Statistics

Table 1 reports the descriptive statistics of the main variables used in the regression analysis, based on the integrated Kaggle datasets covering five major Chinese cities from 2015 to 2022.

Table 1. Descriptive Statistics of Key Variables

Variable	Mean	Std. Dev.	Min	Max
PM2.5 (μg/m³)	47.32	18.41	21.5	89.6
GDP per capita (10^3RMB)	89.14	34.27	42.3	158.9

Variable	Mean	Std. Dev.	Min	Max
Industrial output share (%)	38.56	9.12	22.4	55.7
Energy consumption intensity	0.87	0.21	0.45	1.32
Wind speed (m/s)	2.41	0.73	1.2	4.1

Statistical data show that there are significant differences in PM2.5 concentrations among different cities and in the same city at different times. Cities with more industrial activities and energy consumption usually have higher average pollution levels, while cities with stronger winds usually have lower PM2.5 concentrations.

3.2 Temporal Trends of PM2.5 Concentrations

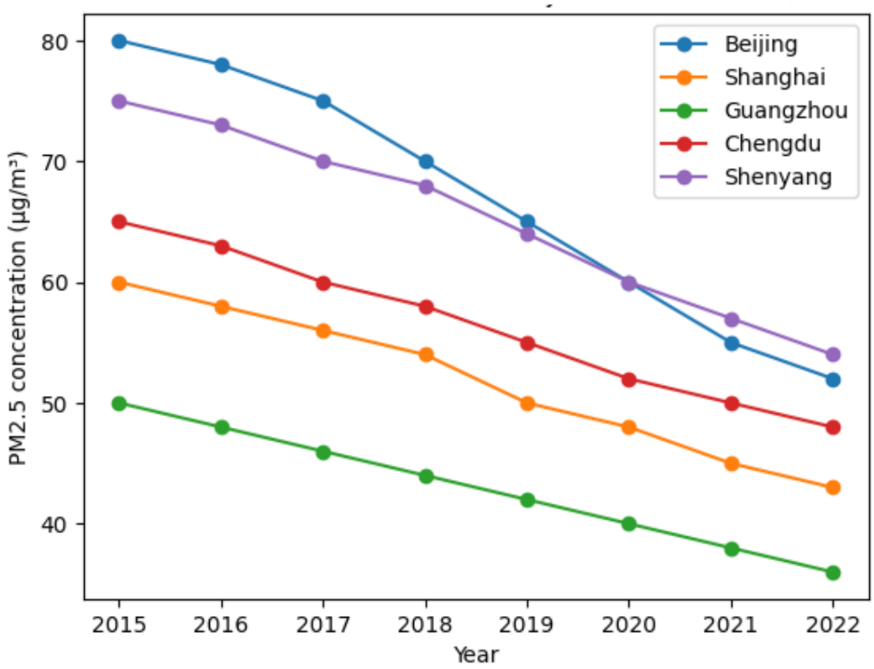

Fig. 1. Annual PM2.5 Concentration Trends in Major Chinese Cities (2015–2022) (Picture credit: Original)

Figure 1 illustrates the annual average PM2.5 concentration trends for the five sampled cities. According to the statistics of PM2.5 in different cities and regions at various times, there are significant differences. Cities with heavier pollution because of more severe pollution may have relatively lower average PM2.5 concentrations, which could be caused by industrial structures; However, coastal areas or areas that are wind-eroded are not as heavily polluted as others.

3.3 Regression Results and Analysis

Table 2 presents the estimation results of the multiple linear regression model. According to the regression results, both GDP per capita and the share of industrial output as well as energy consumption intensity are significantly positively related to PM2.5 concentration. It can be known from this that the expansion of the economy and industry continue to be significant sources of air pollution in cities, which is in line with earlier research results.

Table 2. Multiple Linear Regression Results

Variable	Coefficient	Std. Error	t-value
GDP per capita	0.213***	0.072	2.96
Industrial output share	0.387***	0.094	4.12
Energy consumption intensity	0.451***	0.131	3.44
Wind speed	-0.528***	0.157	-3.36
Constant	12.84	5.21	2.46

Note: *** indicates significance at the 1% level.

The wind speed has a strong negative coefficient; that is to say, a greater atmospheric dispersion capability can effectively weaken the aggregation of PM2.5, which is consistent with the meteorological explanation for the laws of air pollution dispersion.

3.4 Model Diagnostics and Robustness

The variance inflation factor (VIF) test shows that the VIF value of all explanatory variables is less than 5, so multicollinearity is not serious. Heteroskedasticity test results show that the model has mild heteroskedasticity, so robust standard errors are used. The signs and meanings of the coefficients are relatively stable in different model specifications, indicating that the results are robust.

4 Discussion

Based on the above research, it can be determined that China's urban air pollution is caused by a combination of socio-economic activities and meteorological conditions. The main source of industrial production and fossil fuels is still causing PM2.5 pollution; therefore, structural economic transformation and the use of clean energy sources need to be realised in order to tackle it [3,6].

However, there are still some problems. Firstly, the global regression framework is only capable of obtaining the mean effect and tends to conceal the spatial heterogeneity among cities. Secondly, the potential endogeneity problem in economic growth and pollution levels is not examined here. Third, there may be errors in the data used for measurement due to limitations in publicly available data.

In the subsequent research, for the problems that have been reflected above, it is possible to introduce spatial econometric models, dynamic panel methods or relatively accurate pollution data to enhance the accuracy of the research results. Combining machine-learning methods with classical-regression models is likely to be better at capturing nonlinearities and interactions [8,10].

5 Conclusion

Based on public data available in recent times, this paper uses a multiple linear regression method to explore what factors influence PM2.5 pollution in megacities across China. The finding demonstrates that all three variables are positively correlated, as economic development and energy consumption increase, industrial activities will rise accordingly, thereby driving up PM2.5 content; Combined with other factors such as improved weather conditions, a small amount of PM2.5 can be dispersed to the outside world, thus reducing pollution intensity.

In terms of policy, reducing industrial emissions, promoting the use of clean energy, and considering local meteorological conditions are necessary to achieve sustained improvement in urban air quality. In the future, people can improve the accuracy of the analysis with more complete data sets; adopt fine space model methods; combine knowledge from different disciplines. Therefore, it is expected that more effective air quality management and a long-term governance goal will be achieved.

References

[1] Jacob, D. J., Winner, D. A.: Effect of climate change on air quality. Atmospheric Environment, (2009)

[2] Liu, H., Li, X. M., Tian, S. Z., et al.: Spatial-temporal evolution of urban air quality in China and its influencing factors. Ecological Economy, (2021)

[3] Guan, D., Su, X., Zhang, Q., et al.: The socioeconomic drivers of China's primary PM2.5 emissions. Environmental Science & Technology, (2014)

[4] State Council of the People's Republic of China.: Three-Year Action Plan for Winning the Blue Sky Defense War, (2018)

[5] Chen, Z., Kahn, M. E., Liu, Y., et al.: Spatially differentiated air pollution regulation in China. Journal of Environmental Economics and Management, (2020)

[6] Tai, A. P. K., Mickley, L. J., Jacob, D. J.: PM2.5 trends in China: separating emissions and meteorology. Atmospheric Chemistry and Physics, (2019)

[7] Zhang, X., Lin, M.: Socio-economic determinants of urban air pollution in China. Journal of the University of Chinese Academy of Sciences, (2020)

[8] Zhang, Y., Xu, L., Lu, Z.: Synergistic effects influencing urban air quality in China. International Journal of Environmental Science and Technology, (2023)

[9] Wooldridge, J. M.: Introductory Econometrics: A Modern Approach. 7th ed., (2020)

[10] Wei, Y., Wang, Y., Di, Q., et al.: Short-term exposure to PM2.5 and mortality in China. The Lancet Planetary Health, (2021)