Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2025.23143

Exploring Air Quality in Harbin and Zhengzhou by Linear Regression Approach

Peixi Song[1], Yayuan Xiao[2*]

[1]Wuhan Britain-China School, Wuhan, China
[2]Branksome Hall Asia, Jeju, South Korea
**Corresponding Author. Email: xiaoyayuan06091@branksome.asia*

Abstract: In recent years, straw burning has become a major source of air pollution, especially in cities like Harbin and Zhengzhou, where seasonal burning leads to spikes in PM2.5 levels. This study looks into how linear regression can be used to assess air quality, focusing on its ability to measure relationships between different factors and predict pollution levels. By collecting real-world pollution data and factoring in weather conditions, this research applies multiple linear regression to explore how different variables interact and improve prediction accuracy through statistical validation. The results show that this method works better than traditional approaches, offering more reliable estimates. In Zhengzhou, time-weighted regression performed even better, with an R^2 of 97.78%, proving its strength in tracking pollution trends over time. This approach not only makes predictions more precise but also provides solid data to guide environmental policies and pollution control strategies. The study highlights linear regression as a useful tool for air quality analysis and lays the groundwork for future machine learning applications.

Keywords: Air quality, Linear regression, Straw burning.

1. Introduction

As global air pollution continues to worsen, its negative impacts on human health, ecosystems, and daily life have become increasingly evident. Pollutants such as fine particulate matter (PM2.5) and nitrogen oxides pose severe risks, contributing to respiratory diseases, cardiovascular problems, and overall environmental degradation [1]. Consequently, air quality management has emerged as a pressing global concern, necessitating accurate prediction methods for pollutant levels to support effective environmental policies and urban planning [2]. Understanding pollution dynamics through data-driven models is essential for mitigating its adverse effects and ensuring sustainable development [3].

In recent years, numerous studies have explored statistical approaches to air quality prediction, with linear regression emerging as a widely used method. Research has demonstrated that linear regression models can effectively analyze the relationship between meteorological factors—such as temperature, humidity, and wind speed—and pollutant concentrations [4]. More advanced studies have extended this approach by incorporating multiple linear regression techniques, which account for additional variables such as industrial emissions and traffic density [1]. These models have proven useful in identifying key factors influencing air pollution levels and providing a foundation for predictive analysis in urban environments [3].

DOI: 10.54254/2753-8818/2025.23143

This paper explores the application of linear regression in air quality analysis, emphasis its role in identifying key contributing factors and predicting pollutant levels. By creating statistical modelling techniques, the study aims to prove how linear regression can provide valuable insights into air pollution trends and inform environmental policy decisions. A multiple linear regression model is developed using real-world air quality data, illustrating the effectiveness of this approach in quantifying relationships between variables. The dataset used includes pollution records from urban and industrial areas, allowing for a comparative analysis of different pollution sources. Results from the model highlight key correlations between atmospheric conditions and pollution intensity, demonstrating the predictive capabilities of linear regression in environmental research [4].

Additionally, the paper discusses the limitations of linear regression, particularly its assumptions of linearity and independence among predictors. Issues such as multicollinearity and the inability to capture nonlinear relationships are critically assessed, along with potential solutions, including integrating more advanced machine learning techniques. The study also explores how hybrid models that combine linear regression with other statistical or machine learning methods can enhance prediction accuracy. By addressing these challenges, the paper provides a more balanced perspective on the utility and constraints of linear regression in complex environmental systems.

By presenting a comprehensive analysis of linear regression in air quality research, this paper aims to contribute to the broader discussion on data-driven environmental policy. The findings underscore the importance of statistical modelling in understanding pollution dynamics and highlight the need for interdisciplinary approaches to tackling air quality challenges. This research not only emphasizes the practical applications of linear regression but also encourages further exploration of advanced analytical methods to improve air pollution predictions. Ultimately, the study seeks to bridge the gap between theoretical modelling and real-world environmental decision-making, offering valuable insights for researchers, policymakers, and urban planners.

2. Method and theory

Linear regression is one of the most fundamental and widely used techniques in data analysis, statistics, and machine learning. At its core, it aims to understand the relationship between two or more variables and leverage that relationship for predictive modelling. Linear regression seeks to establish a direct correlation between an independent and a dependent variable [5].

Mathematically, linear regression exists in two main forms: simple linear regression and multiple linear regression. The formula for simple linear regression is expressed as follows:

$$y = a + bx + \epsilon \tag{1}$$

where X represents the explanatory variable, Y is the dependent variable, b denotes the slope of the regression line, a is the intercept (the value of Y when X = 0), and ε represents the error term. The most common approach for estimating these coefficients is the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals to achieve the best linear approximation of the data. In cases involving large datasets, alternative optimization techniques such as Gradient Descent can be employed to enhance computational efficiency [6].

The validity of linear regression depends on several critical assumptions. First, linearity assumes a direct, straight-line relationship between independent and dependent variables, which is essential for the model's predictive reliability. Second, independent errors indicate that residuals (the differences between observed and predicted values) should not exhibit correlation, as correlation may lead to biased estimates. Third, homoscedasticity ensures that residuals maintain constant variance across different values of the independent variables; violations of this assumption can lead to inefficient estimators. Fourth, the normality of residuals is crucial for hypothesis testing and confidence interval estimation, as it ensures valid statistical inference. Lastly, in multiple regression,

DOI: 10.54254/2753-8818/2025.23143

the absence of multicollinearity guarantees that independent variables are not highly correlated, since excessive correlation among predictors can produce unstable coefficient estimates and unreliable interpretations. Researchers frequently utilize diagnostic tools such as the Variance Inflation Factor (VIF) to detect multicollinearity, as well as residual analysis techniques to verify model assumptions.

Linear regression finds extensive application across numerous research domains. In environmental science, it helps in analyzing the relationship between air pollution and meteorological factors, as seen in various studies on air quality modelling. In economics, it is widely used to examine how consumer spending correlates with income levels. In the healthcare sector, linear regression models are applied to assess the impact of multiple risk factors on disease prevalence. Marketing analytics also employ linear regression to quantify the effect of advertising expenditure on sales revenue.

Despite its broad applicability, linear regression comes with inherent limitations. If the relationship between variables is nonlinear, a simple linear model may fail to accurately represent the data, necessitating the use of Polynomial Regression or Generalized Linear Models (GLMs). Furthermore, the model is highly sensitive to outliers, as extreme data points can disproportionately influence estimated coefficients and distort analytical results. To mitigate these issues, researchers frequently adopt robust regression techniques or apply data transformations to enhance model reliability.

3. Results and application

3.1. Air quality in Harbin

In Ref. [7], a linear regression method was used to predict the impact of straw burning on PM2.5 concentrations in representative cities. The most important substance affecting urban air quality is PM2.5, and PM2.5 is not only produced by straw burning, but also many other factors. So this article needs to first obtain the background PM2.5 concentration under normal urban conditions. Then subtract the background concentration from the total PM2.5 to obtain the predicted contribution of straw burning to PM2.5.

In this paper, multiple linear regression was used to study concentration estimation in Northeast China. It uses multiple meteorological factor observation data, such as daily average temperature, daily average constant velocity, etc., and air quality detection PM2.5 concentration data. It sings different meteorological molecular data to establish different multiple regression equations. This can be seen from equation

$$Y(x) = a_0 + a_1\mathrm{X}_1 + a_2\mathrm{X}_2 + a_3\mathrm{X}_3 + a_4\mathrm{X}_4 + a_5\mathrm{X}_5 + a_6\mathrm{X}_6 + a_7\mathrm{X}_7 + a_8\mathrm{X}_8 + a_9\mathrm{X}_9 \quad (2)$$

Comparing and selecting several sets of equations with high repeatability, subtracting 1-2 meteorological factors from some combinations. It obtains the minimum error coefficient, fine-tuning and integrating. Finally obtaining the multiple back of daily PM2.5 and meteorological factors were obtained.

Secondly, the meteorological factor data is added to the multiple regression equation. The background concentration fitting result is calculated. Compared with the measured value, the deviation is calculated. The constant with the least difference and the most consistent is found to determine that the final equation can be used. The equation is used to perform a real-life test, and most of the measured deviations are within 10 ug/m^3. It indicates that the multiple regression equation is available. Like Harbin, its multiple regression equation is

$$F(x) = 32 + 0.36\mathrm{X}_1 - 4.49\mathrm{X}_2 + 5.98\mathrm{X}_3 - 0.91\mathrm{X}_5 + 0.89\mathrm{X}_6 - 25.3\mathrm{X}_7 + 12.36\mathrm{X}_8 + 8.2\mathrm{X}_9 \quad (3)$$

Through the multiple regression equation, the measured PM2.5 in several regions can be calculated, and the background concentration value and the contribution of PM2.5 concentration to the impact of straw burning are shown in Table 1. The results in Harbin showed that when the contribution of

DOI: 10.54254/2753-8818/2025.23143

straw burning to PM2.5 value was small. The urban air pollution was relatively low, generally in the middle of good to light pollution. When straw burning emissions contribute significantly to PM2.5 values, they usually reach severe pollution. This method combines meteorological and atmospheric real-time monitoring data. It also uses air quality forecasting statistics to avoid the deviation caused by meteorological data, and further improves the accuracy and operability of prediction.

Table 1: Actual and predicted daily average concentration PM2.5 in Harbin [7]

City	Date	Air quality level	PM2.5 Real value ($\mu g/m^3$)	PM2.5 Background value ($\mu g/m^3$)	Contribution of straw burning impacts on PM2.5 ($\mu g/m^3$)	Background PM2.5 contribution/%	Contribution of straw to burning the measured values $PM_{2.5}$/%
Harbin	3.28	Light	93.0	72.0	21.0	77.4	22.6
	3.29	Light	94.0	74.0	20.0	78.7	21.3
	3.31	Serious	320.0	88.0	232.0	27.5	72.5
	4.1	Severe	151.0	93.0	58.0	61.6	38.4
	4.2	Middle	133.0	113.0	20.0	85.0	15.0
	4.5	Light	106.0	49.0	57.0	46.2	53.8
	4.6	Light	100.0	64.0	36.0	64.0	36.0
Average value			142.4	79.0	63.4	55.5	44.5

3.2. Air quality in Zhengzhou

In Ref. [8], it uses a time-weighted regression approach to analyze Zhengzhou's air quality. The study highlighted that with the continued development of urbanization and industrialization, people are increasingly concerned about ecological and environmental issues, especially air quality. Air quality is critical to public health and environmental sustainability. It makes a focus for research and decision-making.

Start by collecting comprehensive air quality data, including air quality index (AQI) and concentrations of major pollutants, such as PM 2.5, PM 10, sulfur dioxide, and nitrogen dioxide. Also, collect meteorological data that may affect air quality over time. Secondly, the time-weighted regression model was chosen instead of the traditional multiple linear regression [9]. The mathematical formula expression of the TWR model is as follows by

$$y_1 = \beta_0(t_i) + \sum_{k=1}^{p} \beta_k(t_i)\, x_{ik} + \varepsilon_i, i = 1,2 \tag{4}$$

The study showed that the time-weighted regression model provided a better fit to the data, achieving a higher R^2 value of 97.78%, suggesting that it accounted for more AQI variance than the multilinear model. The data was subjected to TWR (Time - Weighted Regression) regression using Arcmap software, and the results are shown in Table 2.

DOI: 10.54254/2753-8818/2025.23143

Table 2: Regression results for Zhengzhou's air quality [8]

Model	Tape width	ALCc	MSE	MAE	R2
TWR	100	7077.79	52.54	5.01	97.78%

In the regression model, time is included as a variable to account for changes in air quality over the selected period. This approach helps capture the dynamic nature of air quality, revealing important trends and seasonal differences in Zhengzhou. The study recognizes that air quality can vary greatly depending on the season, with pollution levels typically higher in winter. The use of multiple regression model prediction can reduce the error and increase the accuracy of prediction.

4. Conclusion

This study begins with an overview of linear regression and a detailed description of its theoretical basis and statistical principles. It is clarified that linear regression aims to find a direct relationship to describe the correlation between independent and dependent variables. Linear regression is then extended to a wide range of research fields, such as economics, marketing analytics, etc., but this article specifically examines its use in air quality in Harbin and Zhengzhou. In addition, the article analyzes in detail the rapid assessment of air quality by straw burning in Harbin, as well as the prediction of air quality in Zhengzhou. It uses real-world air data, such as PM2.5, to list different linear equations for multicollinearity analysis. The equation for the minimum deviation is obtained by calculating and comparing the real data. So as to establish a more accurate multiple linear regression model. Compared with other methods, it is found that the proposed method improves the accuracy of prediction. This article will lay the foundation and standard methodological reference for studying the impact of other pollution sources on air quality. In addition, the authors will continue to conduct research on other urban areas and explore ways to improve air quality.

Authors contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] *Li, P., Li, J., Liu, J., et al. (2025). A method for rapidly assessing the impact of spring straw burning on PM2.5 concentrations in representative cities in Northeast China. China Environmental Monitoring, 1–11.*

[2] *An, M., Liu, E., He, N., Ye, Z., & Hunan Provincial Xiangtan City Meteorological Bureau. (2024). A study on air quality forecasting model in Xiangtan City based on meteorological factors. Sanxia Ecological Environment Monitoring, 4(4), 23–31.*

[3] *Xu, Z., Fang, J., & Chen, Y. (2024). A study on air quality index in Dongguan based on multiple linear regression model. Environmental Science and Technology, 30(2), 40–44.*

[4] *Wang, Z., & Pang, F. (2025). A study on air quality analysis in Zhengzhou City based on time-weighted regression. Henan Science, 1–9.*

[5] *Soffritti, G., & Galimberti, G. (2010). Multivariate linear regression with non-normal errors: A solution based on mixture models. Statistics and Computing, 21(4), 523–536.*

[6] *Sheather, S. J. (2009). A modern approach to regression with R. Springer.*

[7] *Wang, J. J. (2021). A calibration model for air quality monitoring data based on partial least squares regression. Bulletin of Science and Technology, 37(10), 31–37.*

[8] *Chen, X. (2021). Analysis of the influencing factors of air quality index based on linear regression model—A case study of Dazu District, Chongqing. Environmental Impact Assessment, 43(05), 79–82.*

[9] *Zhang, L., & Wang, C. M. (2020). Empirical study on air pollution based on spatial effect multilevel regression model. Journal of Chifeng University (Natural Science Edition), 36(2), 20–24.*