Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260204210

An Interpretability Study of the Air Quality Index in Lanzhou City Based on Random Forest

Guoxuan Zu*

School of Mathematics and Statistics, Lanzhou University, Lanzhou 730000, China

**Corresponding author: Guoxuan Zu.*

Abstract

This study is based on monthly air quality data of Lanzhou City from 2014 to 2022. A Random Forest model is constructed to predict the Air Quality Index (AQI), and interpretability is achieved through feature importance analysis. The results show that PM_{10}, NO_2, and $PM_{2.5}$ are the core factors driving AQI variation, with a cumulative weight of 84.7%. Among them, PM_{10} (46.1%) plays a dominant role, followed by NO_2 (21.9%). The annual average concentration of $PM_{2.5}$ decreased by 42.4% compared with 2014, indicating significant governance effects, yet it remains a key influencing factor. The concentration of NO_2 is higher in winter than in summer, reflecting the impacts of heating and motor vehicles. The model achieves an R^2 of 0.900 on the training set, which decreases to 0.405 on the test set, indicating the presence of overfitting. This study quantitatively identifies the key pollution factors affecting AQI in Lanzhou City, providing a basis for air pollution prevention and control. However, future research should incorporate meteorological factors and improve the model's generalization ability.

Keywords

Lanzhou city, air quality index, random forest, feature importance

1. Introduction

With the advancement of urbanization and industrialization in China, air pollution has become a major challenge affecting public health and sustainable development. As a key indicator for comprehensively evaluating air pollution conditions, the accurate prediction and causal analysis of the Air Quality Index (AQI) are of great significance for pollution early warning and precise control. Studies in regions such as Beijing–Tianjin–Hebei have revealed that the formation of haze pollution in China exhibits high complexity and strong regional characteristics [1]. Therefore, research on the evolution of air quality should be grounded in the fundamental theories of *Air Pollution Control Engineering*, including a thorough understanding of pollution sources, transport and transformation mechanisms, and evaluation standards [2]. Traditional statistical methods have limitations in capturing the complex nonlinear characteristics of air pollution, whereas machine learning methods demonstrate significant advantages due to their strong data-fitting capabilities [3]. Meanwhile, models such as Random Forest have also been applied to analyze the influencing factors of $PM_{2.5}$ concentrations at the national scale [4].

Through methods such as feature importance ranking and SHAP (SHapley Additive exPlanations) values, key factors influencing air quality can be identified and their contributions quantified [5]. Xia et al. [4] used a Random Forest model to interpret the influencing factors of $PM_{2.5}$ concentrations at the national scale. Wang quantitatively assessed the specific impacts of various emission sources and meteorological factors on $PM_{2.5}$ in Lanzhou City by integrating source apportionment models with an interpretable machine learning framework [6]. In addition, rigorous statistical evaluation is another cornerstone for ensuring the reliability of conclusions.

As a valley-basin industrial city, Lanzhou is characterized by conditions under which pollutants tend to accumulate rather than disperse, making its air quality evolution mechanisms of unique research value. Existing studies mainly focus on pollution prediction, while systematic research that integrates interpretability analysis with statistical evaluation and specifically targets Lanzhou City remains insufficient. Therefore, this study takes Lanzhou City as a case, constructs a machine learning framework integrating interpretability analysis and statistical evaluation, and utilizes historical monitoring data to explore the contributions and mechanisms of core pollutants affecting AQI. Meanwhile, the robustness and uncertainty of the model are evaluated, aiming to provide new insights into the causes of air pollution in Lanzhou City and a scientific basis for precise pollution control.

2. Methods

2.1 Data Source Description

The data used in this study were obtained from the Kaggle platform, covering monthly air quality monitoring data of Lanzhou City from January 2014 to December 2022 [7]. As an internationally renowned data science community, datasets released on this platform are typically preliminarily integrated and validated by the community, offering good traceability and providing a continuous and standardized basis for time-series analysis. After acquiring the initial dataset, a rigorous data cleaning process was conducted. All records with missing values in key fields such as “range” and “quality level” were removed. Observations with statistically abnormal pollutant concentrations or those inconsistent with physical meaning were excluded. The “time” variable was standardized and decomposed into more analytically meaningful numerical features such as “year” and “month.” Ultimately, a clean dataset containing 100 valid observations was obtained for subsequent modeling and analysis.

2.2 Indicator Selection and Description

The core objective of this study is to predict AQI using a machine learning model and interpret its driving factors. Therefore, AQI is selected as the target variable of the model, and a feature variable system encompassing both direct pollutant concentrations and derived information is constructed, as shown in Table 1.

Table 1: Description of Variables Used in the Study

Feature Variable	Environmental Significance and Role	Processing and Role in This Study
AQI	A comprehensive indicator integrating multiple pollutant concentrations to quantitatively evaluate air pollution levels and health risks	The ultimate target variable for model prediction and interpretation
Month (m)	Reflects seasonal and periodic patterns of pollution	Transformed into a continuous feature using $\sin\frac{\pi*m}{6}$, serving as a key input to help the model capture cyclic relationships between months
$PM_{2.5}$	The most harmful particulate matter to human health and the main cause of haze	Used to explain its contribution to AQI
PM_{10}	Mainly originates from dust and industrial emissions	Reflects particulate pollution
CO	Produced by incomplete combustion of fuels, indicating local combustion intensity	Serves as an auxiliary indicator of combustion sources
SO_2	Mainly from coal combustion, a marker of industrial emissions	Indicates the impact of industry and energy structure
NO_2	Mainly from motor vehicles and power plants, a precursor of photochemical pollution	Indicates the impact of traffic-related pollution

O_3	A secondary pollutant formed by photochemical reactions of precursors	Explains pollution in summer and autumn

2.3 Method Description

Based on the available dataset, this study utilizes monthly air quality data of Lanzhou City from January 2014 to December 2022, with AQI as the prediction target. The concentrations of six pollutants—$PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, and O_3—along with the month are selected as core feature variables. During the data preprocessing stage, missing values in the "quality level" field were retrospectively imputed based on AQI values, and outlier detection and treatment were conducted for all pollutant concentrations to ensure the quality and stability of the input data.

To capture the periodic variation of air quality across seasons, the "month" variable was specifically processed. Instead of directly using discrete values from 1 to 12, it was transformed into a continuous cyclical feature using a sine function ($\sin(\pi \cdot m/6)$). This transformation enables the model to recognize the cyclic relationship between December and January, thereby effectively learning seasonal patterns such as intensified particulate pollution during the winter heating period and elevated ozone concentrations caused by photochemical reactions in summer. The six pollutant concentration variables, together with the cyclical month feature, constitute the complete set of model input features.

Considering the characteristics of the data, this study adopts the Random Forest regression algorithm. Based on an ensemble of decision trees, this algorithm can flexibly capture complex nonlinear relationships between pollutants and AQI, exhibits strong robustness to outliers and noise, and provides built-in feature importance measures that offer direct and reliable quantitative support for subsequent interpretability analysis. In this study, the first 80 observations were used as the training set, and the remaining 20 observations as the test set. During training, a combination of grid search and cross-validation was employed to optimize key hyperparameters such as the number of trees and maximum depth, in order to balance model fitting capability and generalization performance.

After model training, interpretability analysis focuses on feature importance ranking. Based on the reduction in mean squared error achieved at each split in the Random Forest model, the average contribution of each feature throughout the modeling process is calculated and ranked in descending order. This ranking provides an intuitive and quantitative answer to the global question of "which factors are the key drivers of AQI in Lanzhou City." For example, it allows clear identification of the relative importance of $PM_{2.5}$, PM_{10}, O_3, and seasonal cyclical features in the model. Finally, statistical evaluation is conducted on the test set to comprehensively quantify the model's predictive accuracy and goodness of fit.

3. Results and Discussion

3.1 Results

This study used the monthly air quality data of Lanzhou City from 2014 to 2022, including 100 valid records in total. AQI was set as the dependent variable, and month, $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2, and O_3 were set as the independent variables. A Random Forest regression model was built for prediction. The data were randomly split into a training set (80 records) and a test set (20 records) in an 8:2 ratio. The number of decision trees was set to 100, and the node splitting criterion was MSE. The maximum depth of trees was unrestricted, the minimum sample size for node splitting was 2, and the minimum sample size for leaf nodes was 1. Bootstrapping sampling with out-of-bag testing was enabled. To identify the key drivers of AQI in Lanzhou City, the weight values of each feature in the Random Forest model were calculated. Feature weight reflects the importance of each variable in the model's prediction, with the sum of all feature weights equaling 1. The results are shown in Table 2.

From Table 2, we can see that PM_{10} has the highest weight value, at 0.461, accounting for 46.09% of the total weight, making it the most important variable affecting the model's predictions. The weight value of NO_2 is 0.219, making it the second most important, followed by $PM_{2.5}$ with a weight of 0.167, ranking third. The cumulative weight of these three features is 84.70%, indicating that inhalable particulate matter, nitrogen dioxide, and fine particulate matter are the primary pollutants driving AQI variation in Lanzhou City. The

contributions of the other four features are relatively limited. O_3 has a weight of 0.076, month has a weight of 0.032, and CO and SO_2 have weights of 0.023 and 0.022, respectively, indicating that they have marginal contributions to the model prediction.

Table 2: Feature Weight Values in the Random Forest Model

Feature	Weight Value	Proportion (%)
Month	0.032	3.2%
$PM_{2.5}$	0.167	16.7%
PM_{10}	0.461	46.1%
CO	0.023	2.3%
SO_2	0.022	2.2%
NO_2	0.219	21.9%
O_3	0.076	7.6%

To evaluate the model's prediction accuracy and generalization ability, several performance metrics were computed on both the training and test sets. The results are summarized in Table 3.

Table 3: Model Evaluation Results

Metric	Description	Training Set	Testing Set
R^2	Goodness-of-fit, closer to 1 indicates better	0.900	0.405
Mean Absolute Error (MAE)	Average absolute deviation between observed and predicted values, lower is better	2.912	10.852
Mean Squared Error (MSE)	Mean squared error, closer to 0 is better	44.884	239.128
Root Mean Squared Error (RMSE)	Square root of MSE, same units as original data	6.700	15.464
Median Absolute Deviation (MAD)	Median of absolute residuals, lower is better	1.230	6.285
Mean Absolute Percentage Error (MAPE)	Mean of percentage errors, lower is better	0.023	0.029
Explained Variance Score (EVS)	Proportion of variance explained by the model, ranges [0,1], higher is better	0.903	0.411
Mean Squared Logarithmic Error (MSLE)	Penalizes under-predictions more than over-predictions when RMSE is the same	0.003	0.035

The R^2 value on the training set is 0.900, RMSE is 6.700, and MAE is 2.912, indicating excellent fitting ability on the training data. However, on the test set, the R^2 value drops to 0.405, RMSE rises to 15.464, and MAE increases to 10.852. There is a clear performance gap between the training and test sets, with an R^2 difference of 0.495 and an over one-fold increase in RMSE, suggesting overfitting. The EVS on the test set is 0.411, indicating that the model's ability to explain data fluctuations in unseen data is reasonable but has significant room for improvement. The MAPE on the test set is 0.029, which is within an acceptable range.

The monthly analysis of NO_2 concentrations in 2015, 2017, 2019, and 2021 is shown in Figure 1. From the multi-year average, NO_2 concentrations exhibit significant seasonal variation: concentrations are generally higher in winter months and lower in summer months. For representative years, in January 2015, NO_2 concentration was 54 μg/m³, rising to 73 μg/m³ in December; in January 2017, it was 72 μg/m³, rising to 90 μg/m³ in December; in January 2019, it was 67 μg/m³, rising to 76 μg/m³ in December; in January 2021, it was 64 μg/m³, rising to 72 μg/m³ in December. Overall, NO_2 concentrations were higher during the winter heating period compared to the summer, which is related to increased coal combustion emissions during heating and unfavorable atmospheric dispersion conditions in winter. NO_2's contribution to the Random Forest model ranks second (weight 21.90%), and its significant seasonal fluctuation characteristics provide important information for capturing AQI variations.

Based on the data shown in Figure 2, the annual average concentration of $PM_{2.5}$ exhibits an overall fluctuating downward trend. Specifically, it was 57.33 μg/m³ in 2014, decreased to 50.08 μg/m³ in 2015, and slightly rebounded to 53.50 μg/m³ in 2016. Thereafter, it continued to decline to 33.17 μg/m³ in 2018, followed by slight fluctuations during 2019–2020. In 2021, it reached the lowest value of 28.67 μg/m³ during the study

period, and then slightly increased to 33.00 μg/m³ in 2022. Compared with 2014, the annual average concentration of $PM_{2.5}$ in 2022 decreased by 42.4%, reflecting the positive effectiveness of recent air pollution control measures in Lanzhou City. As the third most important feature in the Random Forest model (weight 16.70%), the long-term declining trend of $PM_{2.5}$ is consistent with the overall improvement in AQI.

Figure 1: Monthly distribution of NO_2 concentrations in different years

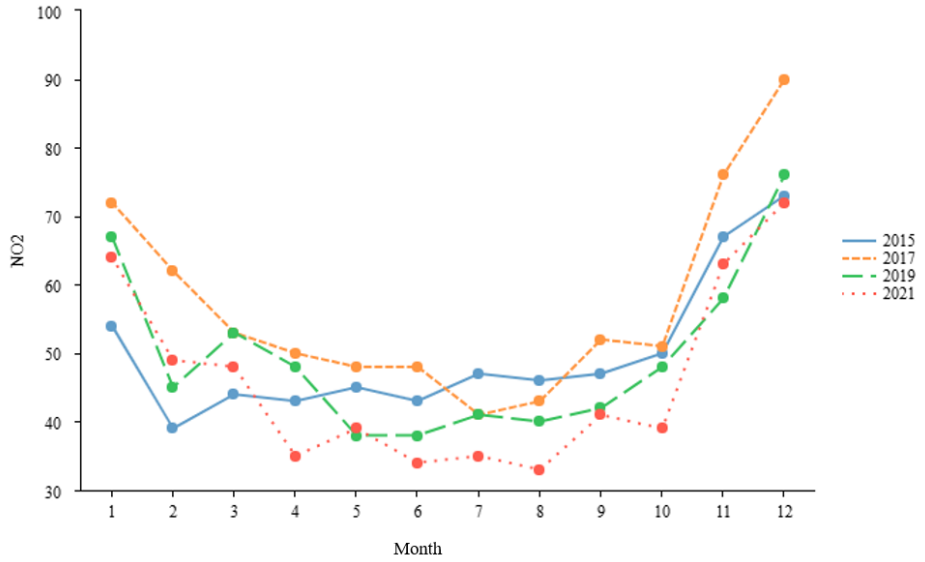
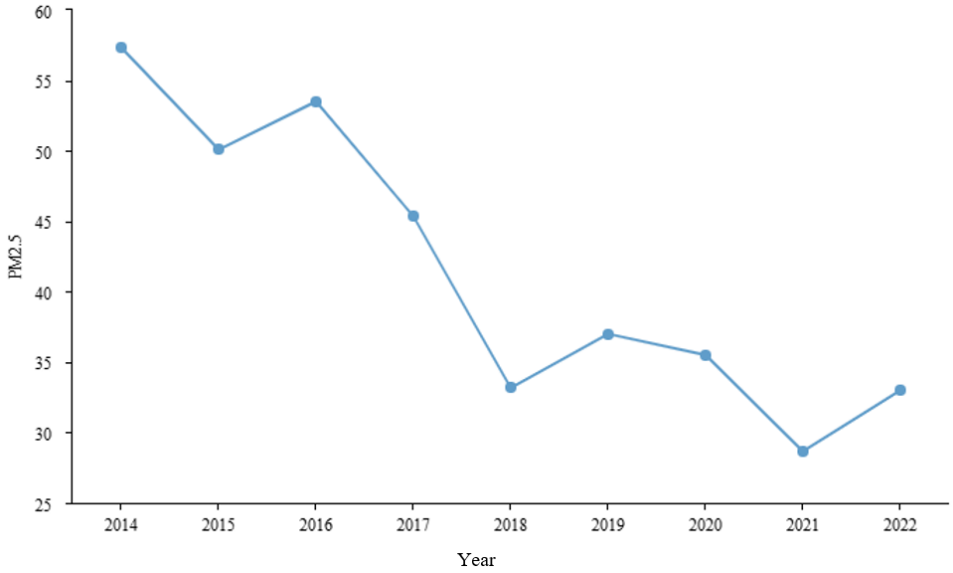


Figure 2: Annual average $PM_{2.5}$ concentration in Lanzhou City from 2014 to 2022



3.2 Discussion

This section discusses the applicability and potential limitations of this study. First, the study is subject to limitations due to the relatively low temporal resolution of the data. Monthly average data smooth out daily fluctuations and short-term pollution processes, making it difficult to capture the rapid evolution of severe pollution events. Second, the feature variables in the model are not sufficiently comprehensive. Only pollutant concentrations and the month feature are included, while meteorological factors and emission source information are absent, which may result in incomplete model interpretation [6]. Third, the model has limited generalization ability. The sample size is relatively small, and the random data split disrupts temporal continuity. The test set R^2 value of 0.405 indicates that the Random Forest model suffers from overfitting. Finally, this study mainly employs feature importance ranking for global interpretability, without conducting local interpretability analysis. As a result, the depth of interpretability is limited, and it is difficult to attribute

prediction deviations in extreme pollution events or specific months. Future research should incorporate daily-scale monitoring data, integrate meteorological factors, adopt time-series cross-validation methods, and introduce local interpretability tools such as SHAP to further analyze individual cases. These improvements would enhance both the interpretability depth and practical applicability of the model [5].

4. Conclusion

In summary, PM_{10}, NO_2, and $PM_{2.5}$ are the core pollutants driving AQI variation in Lanzhou City, with a cumulative feature weight of 84.70%. Among them, PM_{10} ranks first with a weight of 46.1%, indicating that dust pollution is the dominant factor affecting air quality in Lanzhou City. NO_2 ranks second with a weight of 21.9%, and its concentration is significantly higher in winter than in summer, confirming the combined effects of coal combustion for heating and motor vehicle emissions. $PM_{2.5}$ ranks third with a weight of 16.7%, and its annual average concentration has decreased by 42.4% compared with 2014, indicating the positive outcomes of recent air pollution control efforts. However, particulate matter pollution remains a key issue. Model evaluation results show that the R^2 value reaches 0.900 on the training set but decreases to 0.405 on the test set, indicating the presence of overfitting. This suggests that model complexity should be carefully controlled under conditions of limited sample size.

The main contribution of this study lies in introducing interpretability analysis into air quality research in Lanzhou City. By applying feature importance ranking, the study quantitatively identifies the key pollution factors influencing AQI, addressing the limitation of traditional "black-box" models that lack interpretability. The findings provide a scientific basis for precise air pollution control in Lanzhou City: PM_{10}, NO_2, and $PM_{2.5}$ should be prioritized for control; NO_2 emission reduction should be strengthened during the winter heating period; and dust management should be continuously enhanced. Meanwhile, the analytical framework proposed in this study can serve as a methodological reference for understanding air pollution mechanisms in other cities.

References

[1] Wang, Y. S., Zhang, J. K., Wang, L. L., Hu, B., Tang, G. Q., Liu, Z. R., … Ji, D. S. (2014). Significance, current status, and prospects of atmospheric haze pollution research in the Beijing–Tianjin–Hebei region. *Advances in Earth Science*, 29(3), 388–396.

[2] Anonymous. (2021). *Air Pollution Control Engineering* (4th ed.). *China University Teaching*, (5), 98.

[3] Chen, J. C., Dilinuer, Y., Wang, T. Y., Wang, J. Y., Sun, C. X., Xie, X. S., & Feng, W. (2022). Forecasting air pollutant concentrations in Changsha based on machine learning. *Environmental Protection Science*, 48(4), 103–112. https://doi.org/10.16803/j.cnki.issn.1004-6216.2022.04.017

[4] Xia, X. S., Chen, J. J., Wang, J. J., & Cheng, X. F. (2020). Analysis of influencing factors of $PM_{2.5}$ concentrations in China based on a Random Forest model. *Environmental Science*, 41(5), 2057–2065. https://doi.org/10.13227/j.hjkx.201910126

[5] Dong, J. Q., Hu, D. M., Yan, Y. L., Peng, L., Zhang, P. H., Niu, Y. Y., & Duan, X. L. (2023). Identifying driving factors of urban O_3 based on interpretable machine learning. *Environmental Science*, 44(7), 3660–3668. https://doi.org/10.13227/j.hjkx.202208214

[6] Wang, S. S., Wan, Y. Q., Tong, J. L., Liu, Y. L., Liu, H. T., & Ao, C. J. (2025). Source apportionment and quantitative analysis of $PM_{2.5}$ in the main urban area of Lanzhou based on PMF and XGBoost-SHAP. *Acta Scientiae Circumstantiae*, 45(4), 313–321. https://doi.org/10.13671/j.hjkxxb.2024.0510

[7] Kaggle. (2026, January 4). *Kaggle datasets*. https://www.kaggle.com/

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.