

Multi-Station Air Quality Analysis and Prediction in Beijing Based on Deep Learning

Zeyu Zheng^a, Wenjie Hu^a

a. School of Information Engineering, Huzhou University, Huzhou 313000, China

Corresponding author: 2022388513@stu.zjhu.edu.cn

ABSTRACT

With the increasingly severe problem of air pollution, it poses a serious challenge to human health, economic development, and environmental sustainability. How to effectively address this issue through accurate air quality analysis and prediction has become a focus of attention in recent years. This article is based on air quality data from 12 nationally controlled monitoring stations in Beijing from 2013 to 2017. In terms of space, the Air Quality Index (AQI) decreases with increasing distance from the city center, showing a trend of low in the north and high in the south. In terms of time, the peak of AQI occurs at 21:00 and the lowest value occurs at 7:00, which is closely related to people's daily activities. In summer, AQI is relatively good and more stable, while in winter, due to the use of heating equipment and fireworks during the Spring Festival, AQI is poor and fluctuates greatly. Through Pearson correlation analysis, PM_{2.5} and PM₁₀ are the main pollutants that affect AQI, and the impact of meteorological conditions on AQI is relatively small. Although it cannot be immediately manifested, it cannot be ignored in the long-term trend. Finally, Utilizing the Long Short-Term Memory (LSTM) neural network model from deep learning to predict future AQI, with an Mean absolute error (MAE) of 0.7787 and a root mean square error (RMSE) of 0.9719 for 1 to 3 days, indicating high prediction accuracy. This study provides a scientific basis for government air quality management and citizens' daily lives.

Keywords: Air quality index; Spatiotemporal analysis; LSTM.



1. INTRODUCTION

With urbanization and industrialization, air pollution in many countries and cities is becoming increasingly severe. According to the World Health Organization (WHO), air pollution is one of the leading causes of death worldwide, resulting in 7 million deaths annually^[0]. The principal pollutants causing air pollution are carbon monoxide (CO), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), ozone (O₃), and particulate matter (PM_{2.5} and PM₁₀). Over 90% of people breathe air in which the pollutant content exceeds the levels recommended by the World Health Organization. People in emerging countries are most severely affected, particularly in China, a country with a population of 1.4 billion^[2]. In recent years, the impact of air pollution on health has been widely studied. Previous studies have shown that long-term inhalation of air pollutants can have adverse effects on human health. Air pollution has both acute and chronic effects on human health, affecting many different systems and organs, and may lead to an increase in mortality or serious diseases, or may pose existing or potential harm to human health^[3]. Long-term exposure to pollutants such as atmospheric particulate matter and ozone is associated with increased mortality and hospitalization from respiratory and cardiovascular diseases. Air pollution has also been linked to central nervous system (CNS) diseases, including stroke, Alzheimer's disease, Parkinson's disease, and neurodevelopmental disorders^[4]. Millions of people around the world are chronically exposed to air pollutants at concentrations well above legal safety standards^[5]. Therefore, morbidity and mortality caused by air pollution remain a growing public health concern worldwide. Air pollution ranks eighth among the leading risk factors for death, accounting for 2.5% of all deaths in developed countries^[6]. Epidemiological and observational studies have found strong links between exposure to ambient air pollutants and adverse health outcomes, such as hospitalization and mortality^[7]. Over the past decade, research on Beijing's air quality has gained increasing attention, as the city has long faced severe pollution, leaving a strong impression on residents and visitors. Commonly used pollution parameters like O₃, PM_{2.5}, PM₁₀, and are often used to assess air quality^[8]. However, the index of a single pollutant cannot fully reflect the overall air quality of a region, nor can it accurately measure the impact of air pollution on human health. Therefore, the use of the Air Quality Index (AQI) has gradually become the mainstream trend in assessing air quality. The time change of air quality has important indicative significance, and the change of different time scales can reveal the pollution trend and air quality characteristics in a specific time period. Annual changes reflect long-term trends in air quality improvement or deterioration and help assess the effectiveness of environmental policies and changes in emissions sources such as industry and transportation^[9]. Monthly variations can reveal more detailed short-term trends, identify

pollution peaks in specific months, and help fine-tune pollution prevention measures. Effective air quality analysis and prediction is crucial for reducing the impact of air pollution. It provides authorities with advanced information to prevent hazardous events and design mitigation measures. Early air quality predictions relied on statistical methods like Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA), but these models struggle with feature extraction and generalization, especially with big data. Artificial intelligence (AI) has become popular for air quality prediction due to its ability to learn from non-linear time series data^[10]. In 1993, machine learning models like multi-layer perceptrons (MLP) were first applied to SO₂ concentration prediction^[11], followed by the use of artificial neural networks (ANN) and support vector machines (SVM) for air pollution prediction. Though SVM has higher error rates than ANN, it can still reduce MAE and MSE with larger datasets^[12]. However, as IoT data grows and prediction accuracy demands increase, traditional models and machine learning algorithms face limitations^[13]. Deep learning, especially LSTM networks, has proven effective for big data prediction, addressing gradient disappearance issues in recurrent neural network (RNN) to capture long-term dependencies in time series. LSTM models have outperformed traditional methods in predicting PM_{2.5} levels from various monitoring stations^[14].

Air quality analysis and prediction play a crucial role in reducing air pollution. By accurately assessing the changing trend of air pollutants and the source of pollution, and timely predicting air quality fluctuations, it can provide early warning for the government and the public, take countermeasures in advance, reduce the harm to public health, and promote the continuous improvement of air quality. Realize the dual guarantee of environmental protection and public health. This study aims to provide a scientific basis for new policy recommendations to improve air quality in Beijing.

In summary, a brief review of existing air quality prediction methods was provided, highlighting the limitations of traditional statistical models such as ARMA and ARIMA in handling complex, nonlinear time-series data. The advantages of deep learning methods, particularly LSTM, were emphasized for their ability to capture long-term dependencies and complex patterns. However, existing research still faces challenges related to data diversity, model generalization, and interpretability, especially regarding the integration of spatial, meteorological, and other multidimensional data to enhance prediction accuracy. Finally, future research directions were proposed, suggesting the exploration of hybrid models, the integration of real-time data streams, and the development of explainable AI models to improve both prediction accuracy and adaptability.

2. MATERIALS AND METHODS

2.1. Study area and datasets

Beijing, China's capital, is a political, cultural, and technological hub, but also one of the most polluted cities globally. Its semi-humid, semi-arid climate and rapid urbanization have worsened air pollution, impacting residents' quality of life. Studying Beijing's air quality is crucial for improving local health and offering insights for other cities with similar challenges.

The dataset, provided by the University of California, Irvine (UCI), includes hourly air quality data from 12 monitoring stations in Beijing, recorded from March 1, 2013, to February 28, 2017. It covers six major pollutants (CO, NO₂, SO₂, O₃, PM_{2.5}, PM₁₀) and six meteorological variables (temperature, barometric pressure, dew point, precipitation, wind direction, wind speed). The Air Quality Index (AQI) is calculated based on pollutant concentration limits and their effects on human health and the environment, as outlined in China's Ambient Air Quality Index Technical Regulations (HJ633-2012). The AQI value with the highest sub-index is selected, as shown in Table 1 and Equation (1). This dataset provides valuable insights for studying air quality trends and influencing factors, supporting the development of effective air quality management strategies.

Table 1 Air Quality Index and corresponding item concentration limits of pollutants

Indoor Air Quality Index (IAQI)	Pollutant item concentration limit						
	SO ₂	SO ₂	NO ₂	PM ₁₀	CO	O ₃	PM _{2.5}
	(μg/m ³)	(μg/m ³)	(μg/m ³)	(μg/m ³)	(mg/m ³)	(μg/m ³)	(μg/m ³)
	1-hour	24-hour	1-hour	1-hour	1-hour	1-hour	1-hour
0	0	0	0	0	0	0	0
50	150	50	100	50	5	160	35
100	500	150	200	150	10	200	75
150	650	475	700	250	35	300	115
200	800	800	1200	350	60	400	150
300	(1)	1600	2340	420	90	800	250
400	(1)	2100	3090	500	120	1000	350
500	(1)	2620	3840	600	150	1200	500
explanation	(1) When the 1-hour average concentration of SO ₂ is higher than 800μg/m ³ , report the index calculated based on its 24-hour average concentration.						

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} (C_P - BP_{Lo}) + IAQI_{Lo} \quad (1)$$

In the formula, $IAQI_P$ represents the air quality sub-index of pollutant P, C_P is the mass concentration of pollutant P, BP_{Hi} and BP_{Lo} are the high and low concentration limits for

pollutants similar to Cp, respectively, and $IAQI_{Hi}$ and $IAQI_{Lo}$ are the corresponding air quality sub-indices for BP_{Hi} and BP_{Lo} .

2.2. Methods

This study analyzes and predicts data to uncover patterns, with key analyses including: (a) mapping AQI spatial changes in Beijing from 2013 to 2017 and visualizing distribution trends; (b) exploring AQI's temporal changes using boxplots to examine daily and seasonal variations; (c) analyzing the correlation between pollutants, meteorological factors, and AQI to identify the most influential time factors.

We discuss the AQI prediction model using LSTM, a type of RNN designed to handle long-term dependencies. Traditional RNNs struggle with vanishing or exploding gradients, but LSTM overcomes this issue with three gates—forget, input, and output. The forget gate decides what information to discard, the input gate determines what new information to add, and the output gate generates the final output. These gates enable LSTM to preserve long-term correlations, as shown in equations (2) – (7).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t * \tanh(C_t) \quad (7)$$

In the formula, $\tilde{C}_t$ is the candidate cell state, C_t is the current cell state, and C_{t-1} is the previous cell state. h_t and h_{t-1} are the current and previous hidden states, and x_t is the input. f_t , i_t , o_t are the outputs of the forget, input, and output gates. W_f , W_i , W_o , W_c are the weights, and b_f , b_i , b_o , b_c are the biases.

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are common metrics for model evaluation. RMSE measures the deviation between predicted and true values, with smaller values indicating better performance. Its range is $[0, +\infty)$, calculated as the square root of the sum of squared differences, as shown in Equation (8). MAE calculates the average of absolute prediction errors, focusing on error magnitude. Its range is $[0, +\infty)$, with a value of 0 indicating perfect prediction, and is computed as shown in Equation (9).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9)$$

In the formula, n is the number of samples; y_i is the true value; $\hat{y}_i$ is the predicted value; $\bar{y}$ is the average of all true values.

3. RESULTS

3.1. AQI Spatial Variation

Between 2013 and 2017, the AQI spatial distribution at 12 Beijing monitoring stations showed clear regional differences (Figure 1). AQI values decreased with distance from the city center, with Dongcheng and Xicheng districts having the highest AQI values, reaching 118.148 in Shijingshan District. This reflects the impact of traffic, energy consumption, and population density. Urban centers, with higher energy and vehicle density, have worse air quality, while non-central areas with more trees and lower urbanization show better air quality^[15]. Improving Beijing's air quality requires clean energy, better traffic management, and more green spaces, offering key insights for sustainable development.



Figure 1 Spatiotemporal distribution characteristics of AQI

3.2. AQI Time Variation

Hourly analysis (Figure 2) shows the average AQI in Beijing from 2013 to 2017 was 111.35, with a peak of 122.51 at 21:00, likely due to traffic congestion and exhaust emissions during the evening rush. The AQI is lower at 7:00 (101.99) due to reduced human activity and lower wind speeds trapping pollutants overnight, while the reduced pollutant production also contributes to lower AQI values. Seasonal analysis (Figure 3) shows the average AQI in summer (95.04) is lower than the annual average, likely due to rainfall and strong solar radiation aiding in pollutant diffusion^[16]. Winter has the highest AQI (124.82) due to increased heating emissions and fireworks during the Spring Festival, which release a large number of pollutants and worsen air quality^{[17][18]}.

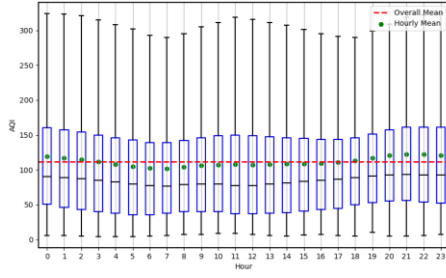


Figure 2 Hour Variation

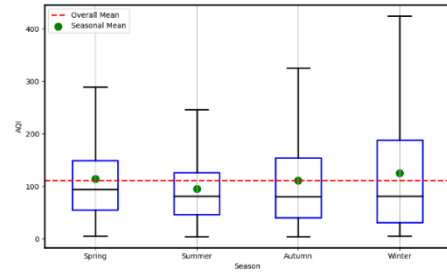


Figure 3 Seasonal Variation

3.3. Analysis Of Influencing Factors

Air quality is mainly influenced by pollutants like CO, NO₂, SO₂, O₃, PM_{2.5}, and PM₁₀, from vehicle exhaust and industrial sources^[19]. Figure 4 shows a strong correlation between AQI and PM_{2.5}/PM₁₀, with their concentrations rapidly worsening AQI. CO and NO₂ also significantly impact AQI, while SO₂, though less correlated, can worsen air quality at high concentrations. O₃, a secondary pollutant, can increase under strong sunlight. Meteorological factors like wind speed, temperature, and precipitation affect pollutant diffusion and accumulation, with wind aiding pollutant dilution and temperature and dew point influencing moisture retention. Wind direction, pressure, and precipitation impact long-term air quality trends^[20]. As shown in Figure 5, these factors have a cumulative effect on air quality.

Time factors, such as daily rhythms, seasonal changes, and special events, greatly impact air quality. Daytime activity typically worsens air quality, while nighttime calm can improve it. Holidays, with reduced traffic and industrial activity, also influence air quality^[21]. Using Random Forest for feature selection and SHAP for feature importance measurement, the study identifies seasonal, hourly, and working day characteristics as key time-related factors contributing to AQI, as shown in Figure 6.

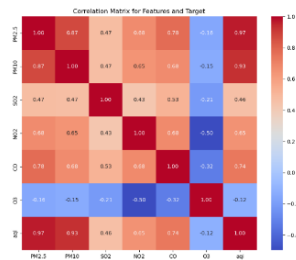


Figure 4 Correlation of pollutant emission

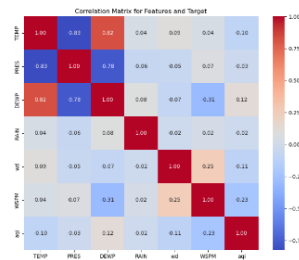


Figure 5 Correlation of meteorological conditions

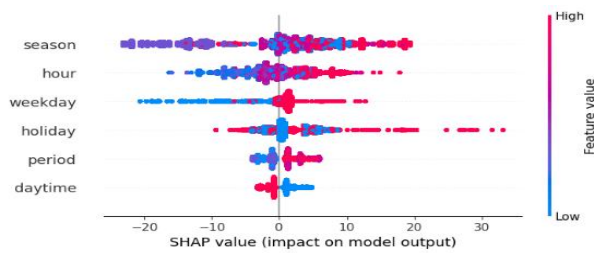


Figure 6 SHAP Summary diagram of time factor

3.4. Prediction

Our experiment uses pollutant, meteorological, and spatiotemporal data. Pollutant data includes hourly concentrations of CO, NO₂, SO₂, O₃, PM2.5 and PM10. Meteorological data includes temperature, pressure, dew point, precipitation, wind direction, and wind speed. Spatial data includes the air quality monitoring station's longitude, latitude, and its distance from Tiananmen, Beijing's center. Three key time features—seasonal, hourly, and working day characteristics—are selected based on their contribution to AQI.

Experimental hyperparameters are shown in Table 2. Using an LSTM model for 1, 2, and 3-day AQI predictions, the results, as shown in Table 3, show a MAE of 0.7787 and RMSE of 0.9719. These results highlight the model's ability to accurately capture AQI dynamics, supporting better air quality management, policy-making, and public health awareness. This predictive model is crucial for improving air quality monitoring and management efficiency.

Table 2. Experimental Hyperparameters

Config	Values
optimizer	Adam
learning rate	0.0001
look-back window	72
prediction length	[24,48,72]
batch size	32
patience	3
epoch	10

Table 3. Prediction Results Table

Model	Metric	24	48	72	Avg
LSTM	MAE	0.7012	0.8156	0.8194	0.7787
	RMSE	0.8953	1.0093	1.0111	0.9719

4. CONCLUSIONS

This study analyzed air quality data from 12 monitoring stations in Beijing from 2013 to 2017. Spatial analysis shows that the AQI decreases with distance from the city center, with Dongcheng and Xicheng districts experiencing higher AQI values due to traffic congestion, high population density, and energy consumption. Non-central areas benefit from higher tree coverage and lower urbanization, improving air quality. Time analysis reveals that AQI peaks around 21:00 during the night rush hour and is lowest at 7:00, reflecting traffic patterns. Seasonal analysis shows the best air quality in summer due to rainfall, convection, and high temperatures, while winter sees a spike in AQI due to heating and fireworks. Pearson correlation analysis indicates that PM2.5 and PM10 are the main pollutants, while CO and NO₂ also impact AQI. O₃ has a negative correlation, suggesting it may improve AQI in some

cases. Although meteorological conditions have a low short-term correlation with AQI, wind speed, precipitation, and pressure regulate air quality in the long term, with wind speed aiding in pollutant diffusion. The LSTM model's AQI prediction for 1, 2, and 3 days showed high accuracy, with MAE of 0.7787 and RMSE of 0.9719, indicating its effectiveness in capturing AQI dynamics and supporting decision-making.

Future research can improve AQI prediction by integrating socio-economic, geographic, and industrial data, and extend time series analysis to explore long-term policy impacts. Additionally, spatial heterogeneity and regional pollution control strategies, including traffic, industry, and greening, should be studied further. Developing an intelligent early warning system based on LSTM could provide real-time air quality forecasts and health advice to the public, enhancing environmental management and public health.

Acknowledgements: The authors are grateful for the supports provided by the Huzhou University.

REFERENCES

- [1] D.A. Vallero. Fundamentals of air pollution, *Academic press*, (2014).
- [2] S. Wang, J. Hao. Air quality management in China: Issues, challenges, and options, *Journal of Environmental Sciences*, 24 (2012) 2-13.
- [3] M. Kampa, E. Castanas. Human health effects of air pollution, *Environmental pollution*, 151 (2008) 362-367.
- [4] S. Genc, Z. Zadeoglulari, S.H. Fuss. The adverse effects of air pollution on the nervous system, *Journal of toxicology*, (2012).
- [5] M.L. Block, L. Calderón-Garcidueñas. Air pollution: mechanisms of neuroinflammation and CNS disease, *Trends in neurosciences*, 32 (2009) 506-516.
- [6] K.M.V. Narayan, M.K. Ali, J.P. Koplan. Global noncommunicable diseases—where worlds meet, *New England Journal of Medicine*, 363 (2010) 1196-1198.
- [7] M. Goldberg. A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases, *Reviews on environmental health*, 23 (2008) 243-298.
- [8] J. An., Y. Wang, B. Zhu. Measurements of O₃, NO_x and VOCs during summer in Beijing, China, *Environmental Engineering and Management Journal*, 15 (2016) 715–724.
- [9] S. Wang, M. Zhao, J. Xing. Quantifying the air pollutants emission re-duction during the

2008 Olympic Games in Beijing, *Environmental Science & Technology*, 44 (2010) 2490–2496.

[10] W. Niu, Z. Feng. Evaluating the performances of several artificial intelligence methods in forecasting daily streamflow time series for sustainable water resources management, *Sustainable Cities and Society*, 64 (2021) 102562.

[11] M. Boznar, M. Lesjak, P. Mlakar. A neural network-based method for short-term predictions of ambient SO₂ concentrations in highly polluted industrial areas of complex terrain, *Atmospheric environment. Part B. urban atmosphere*, 27 (1993) 221-230.

[12] A. Shawabkeh, F. Al-Beqain, Redan A, et al. Benzene air pollution monitoring model using ann and svm, *Fifth HCT Information Technology Trends (ITT)*, (2018) 197-204.

[13] X. Hao, G. Zhang, S. Ma. Deep learning, *International Journal of Semantic Computing*, 10 (2016) 417-439.

[14] Y.S. Chang, H.T. Chiao, S. Abimannan. An LSTM-based aggregated model for air pollution forecasting, *Atmospheric Pollution Research*, 11 (2020) 1451-1463.

[15] Y. Tian, Y. Jiang, Q. Liu. Temporal and spatial trends in air quality in Beijing, *Landscape and urban planning*, 185 (2019) 35-43.

[16] W. Chen, L. Yan, H. Zhao. Seasonal variations of atmospheric pollution and air quality in Beijing, *Atmosphere*, 6 (2015) 1753-1770.

[17] R. Shen, K. Sch äfer, Schnelle-Kreis J, et al. Seasonal variability and source distribution of haze particles from a continuous one-year study in Beijing, *Atmospheric Pollution Research*, 9 (2018) 627-633.

[18] M. Elser, R.J. Huang, R. Wolf. New insights into PM 2.5 chemical composition and sources in two major cities in China during extreme haze events using aerosol mass spectrometry, *Atmospheric Chemistry and Physics*, 16 (2016) 3207-3225.

[19] C. Zhao, M. van Heeswijk, J. Karhunen. Air quality forecasting using neural networks, *Symposium Series on Computational Intelligence (SSCI)*, (2016) 1-7.

[20] A.M. Fiore, V. Naik, D.V. Spracklen. Global air quality and climate, *Chemical Society Reviews*, 41 (2012) 6663-6683.

[21] G. Yun, Y. He, Y. Jiang. PM_{2.5} spatiotemporal evolution and drivers in the Yangtze River Delta between 2005 and 2015, *Atmosphere*, 10 (2019) 55.