Opportunities and Challenges in Sustainability
https://www.acadlore.com/journals/OCS

Spatial and Seasonal Dynamics of Air Quality in the Beijing-Tianjin-Hebei Region: An Analysis Using K-means Clustering and BP Neural Networks

Yuanyuan Wang[1], Zhuang Wu[1*], Wanshu Fu[1], Jiaqi Du[1], Yi Zhang[2]

[1] School of Management and Engineering, Capital University of Economics and Business, 100070 Beijing, China
[2] School of Environment, The University of Manchester, M13 9PL Greater Manchester, UK

[*] Correspondence: Zhuang Wu (wuzhuang@cueb.edu.cn)

Received: 09-25-2023 **Revised:** 10-29-2023 **Accepted:** 11-17-2023

Citation: Wang, Y. Y., Wu, Z., Fu, W. S., Du, J. Q., Zhang, Y. (2023). Spatial and seasonal dynamics of air quality in the Beijing-Tianjin-Hebei region: An analysis using K-means clustering and BP Neural Networks. *Oppor Chall. Sustain.*, *2*(4), 184-196. https://doi.org/10.56578/ocs020402.

Abstract: In the context of rapid economic development, air pollution has emerged as a critical environmental issue, particularly in the Beijing-Tianjin-Hebei region. This study, through the application of Air Quality Index (AQI) data and K-means clustering, investigates the seasonal variations and spatial distribution of air quality in this region. It has been identified that air pollution in this area is not only subject to seasonal fluctuation but also exhibits distinct patterns of local spatial aggregation. Utilizing a Back Propagation (BP) Neural Network model, this research predicts AQI values, offering foresight into the development and transformation of haze weather conditions. The findings of this investigation are instrumental in enhancing the understanding of air pollution dynamics, facilitating the formulation of effective air control strategies. Such strategies are vital for the issuance of accurate pollution warnings and reminders, thereby contributing to the mitigation of severe pollution impacts.

Keywords: Air Quality Index; K-means clustering; Local spatial aggregation; Back Propagation Neural Network model

1. Introduction

The indispensability of clean air for human existence is juxtaposed starkly against the backdrop of industrialization and the rapid pace of economic development, which have culminated in an overabundance of air pollutants. This saturation of pollutants has transformed haze from a natural phenomenon into a predominantly anthropogenic one, posing significant threats to public health, traffic safety, and the overall quality of life. Urban centers, in particular, are grappling with this hazard, and nowhere is this more evident than in China. The nexus of technological transformation and swift economic expansion has exacerbated air pollution, especially in economically pivotal regions such as the Beijing-Tianjin-Hebei and the Yangtze River Delta. Metropolitan hubs including Beijing, Tianjin, Shijiazhuang, Zhengzhou, and Nanjing are at the epicenter of this crisis, with pollution radiating to adjacent cities and precipitating concentrated episodes of contamination. The elucidation of the spatiotemporal dynamics of haze pollution in China in recent years emerges as an imperative for governmental intervention. Such an understanding is foundational for the development of efficacious haze control policies, the enhancement of regional collaborative prevention efforts, and the furtherance of China's ecological civilization initiative.

The ensuing sections of this study are organized as follows: Section 2 presents a review of the relevant literature; Section 3 delineates the research methodology and data sources employed; Section 4 offers an analysis of the results; and Section 5 encapsulates the conclusions drawn from this study.

2. Literature Review

The multifaceted nature of haze formation, its influencing factors, and predictive methodologies has captivated scholarly attention globally. Li (2018) conducted an analysis utilizing cluster analysis to explore the genesis and control measures of haze, revealing a higher concentration of SO_2 and inhalable particulate matter in northern Chinese cities compared to their southern counterparts, thus indicating regional disparities in urban pollutant concentration distribution. Van et al. (2022) provided a comprehensive summary of the characteristics, causative factors, and mechanisms behind haze formation in Southeast Asia. Their findings highlighted the varying trends in haze occurrence frequency and intensity across cities, attributing these differences not only to local pollution sources like biomass burning but also to meteorological factors and long-range transport. The role of secondary aerosols in haze formation was also underscored. Cai et al. (2023) examined the impact of both natural and socio-economic factors on haze pollution in China. The study found positive correlations between haze pollution and variables such as temperature, atmospheric pressure, population density, and green coverage rate in built-up areas, whereas per capita GDP exhibited an inverse relationship. However, the extent of influence varied among these factors.

Susilo & Putranto (2023) investigated various determinants affecting the provincial air quality index in Indonesia from 2012 to 2019. Their research highlighted that circular economy variables, including water resource efficiency, waste treatment, and waste production, significantly improved air quality. However, variables related to the population's contribution to coal and water use efficiency did not demonstrate a positive impact on Indonesia's air quality. Abdul-Rahman et al. (2024) synthesized information on the effects of air quality on cardiovascular health. The study underscored that air pollution poses a global health challenge, elevating the risk of cardiovascular diseases such as heart attacks, strokes, and arrhythmias, with particulate matter, especially $PM_{2.5}$ and ultrafine particles, being pivotal in the adverse effects of air pollution on cardiovascular health. Xu et al. (2023) explored whether the construction of low-carbon cities can reduce pollution, and the research results showed that the construction of low-carbon cities did not effectively reduce haze pollution in pilot cities. Further research has found that the failure of the emission reduction effect of haze pollution may lie in the failure of the technological innovation effect and population quality effect of low-carbon city construction, as well as regional heterogeneity. Supphapipat et al. (2023) delved into the effects of air pollution on post-organ transplant outcomes, asserting that air pollution creates a hazardous environment affecting not only global human health but also post-transplant outcomes.

Ma & Cao (2021) investigated the impact of haze pollution on industrial structure, revealing a significant positive spatial correlation in haze pollution in China that remains relatively stable. The study suggested that the optimization and rationalization of industrial structure, technological progress, and trade opening are effective in reducing haze pollution, with market mechanisms proving more efficacious than governmental interventions. Wang & Xu (2022) studied the impact of digitalization on haze pollution and explored the mediating role of energy consumption. Empirical results showed that digitalization can effectively suppress haze pollution, and this inhibitory effect has significant heterogeneity. In addition, digitalization can indirectly suppress haze pollution by reducing energy consumption intensity and optimizing energy consumption structure. Jia & Yan (2022) studied the impact of haze pollution on the demand for commercial health insurance. The results showed that the relationship can show significant regional heterogeneity, with a significant positive correlation in the eastern region and a significant negative correlation in the central and western regions. Lv et al. (2022) explored the impact of government attention to the environment on haze pollution, and the results showed that local government environmental attention effectively reduced haze pollution. After considering robustness, the conclusion still holds. Liu et al. (2023a) studied how different types of industrial agglomeration contribute to haze pollution. The results show that there is an inverted U-shaped relationship between related variety and haze pollution, however, the overall variety aggravates haze pollution. Wu (2023) assessed the efficacy of joint prevention and control measures for air pollution in China, concluding that such measures significantly influence both the overall and individual pollutant emissions. Li et al. (2023) examined the interplay between market segmentation and haze pollution in the urban agglomerations of China's Yangtze River Delta, finding that cities with high market segmentation and haze pollution could potentially reduce pollution through future market integration. Many scholars have predicted AQI by constructing models, such as Liu & Guo (2022) predicting the Air Quality Index (AQI) based on LSTM model and SSA algorithm, Zhang et al. (2022) predicting AQI based on real-time images of deep learning, Duangsuwan et al. (2022) conducted AQI mapping and data evaluation based on real-time air pollution monitoring using low altitude drones. Liu et al. (2023b) developed an AQI prediction model based on a BP Neural Network, providing a reliable reference for governmental decision-making. Lastly, Ahmad & Ahmad (2023) studied an AQI prediction model based on a layer-recurrent neural network, demonstrating its utility for the Air Pollution Control Bureau in enhancing the accuracy of AQI predictions and pollution control measures. Si et al. (2023) studied a new algorithm for haze recognition based on FY3D/MERSI-II remote sensing data.

While the existing body of research lays a solid scientific groundwork for the formulation of haze prevention and control strategies, it is not without its limitations. A prevalent issue observed is the focus on either socio-economic or natural meteorological factors in isolation, leading to a fragmented understanding of haze formation mechanisms and treatment methodologies. Moreover, a majority of these studies have centered their analysis on $PM_{2.5}$ and PM_{10} as primary haze pollution indicators, overlooking the fact that different cities may have varied primary pollutants contributing to pollution. This reliance on singular indicators fails to accurately encapsulate the complete pollution scenario, often resulting in the underestimation of actual pollution levels in certain urban areas.

In addressing these gaps, this study advocates for the utilization of AQI as a more comprehensive measure of air pollution. The AQI, a dimensionless index, quantitatively represents air quality by amalgamating six indicators: PM_{10}, $PM_{2.5}$, sulfur dioxide, nitrogen dioxide, ozone, and carbon monoxide. An escalated AQI value is indicative of deteriorating air quality and an increased risk to public health. Furthermore, this research incorporates visual software tools to analyze the mass concentration, as well as the spatial and temporal distribution of pollutants, placing emphasis on regional disparities and interdependencies. A BP Neural Network model, developed in Matlab, facilitates the prediction of AQI time series. This approach enables the systematic identification and quantification of both natural and anthropogenic factors influencing air quality, alongside their spatial spillover effects. The implications of this study are profound, addressing a pressing need for the formulation of globally coordinated development plans. It holds substantial scientific relevance for the establishment of regional and urban atmospheric environment prevention and control policies, thereby contributing to a holistic understanding and management of air pollution dynamics.

3. Research Methodology

In the contemporary era of vast data proliferation, the effective analysis and interpretation of this data is paramount, yet it often presents a challenge due to its sheer volume and complexity. To address this, data visualization technology has emerged as a pivotal tool. It facilitates the transformation of complex datasets into comprehensible visual formats such as charts, graphs, and other aids. This technology not only streamlines the data analysis process but also unravels underlying patterns and insights, thereby enhancing decision-making processes.

In this study, GeoDa, a sophisticated geographic mapping software, was employed to process and evaluate air quality monitoring data from 13 cities in the North China Plain. The software generated a map that visually represented varying pollution levels across these cities using a color-coded scheme. Additionally, a tabular format was used to present the concentrations of various air factors in the Beijing-Tianjin-Hebei region. During the development of the BP Neural Network model, the data underwent a normalization process to ensure accuracy and consistency in the model's predictive capabilities.

3.1 Air Quality Evaluation Method (AQI)

AQI is utilized as a comprehensive measure for assessing air pollution levels and understanding their potential health impacts. Prior to the adoption of AQI, the Air Pollution Index (API), which included only three pollution indicators, was employed in China. In response to the evolving complexity of air pollution, China integrated foreign expertise and methodologies, tailored to its specific environmental context, to develop and adopt the AQI. This initiative has significantly enhanced China's capabilities in addressing the multifaceted challenges of air pollution.

In this study, pollution is the primary focus, with an emphasis on analyzing the temporal variation characteristics of different pollutants and assessing their compliance with established standards. The AQI method is employed to perform an exhaustive evaluation of air quality. This approach effectively underscores the impact of individual pollutants on overall air quality. Specifically, the Individual Air Quality Index (IAQI), correlating to the concentration of a particular pollutant, is instrumental in determining air quality. The sub-index I_P for air quality, associated with the mass concentration C_p of a specific pollutant P, is calculated using the following Eq. (1):

$$I_P = \frac{I_{ph} - I_{pl}}{C_{ph} - C_{pl}}\left(C_p - C_{pl}\right) + I_{pl} \tag{1}$$

where, C_{ph} and C_{pl} represent the highest and lowest values of the concentration limits for a similar pollutant, respectively, and I_{ph} and I_{pl} correspond to the IAQI values for C_{ph} and C_{pl}.

$$\text{AQI} = \max\left\{IAQI_1, IAQI_2, IAQI_3, \ldots IAQI_n\right\} \tag{2}$$

The AQI value is instrumental in identifying the primary pollutant when the AQI exceeds 50, as it is the pollutant with the largest IAQI. The evaluation of air quality levels and categories is conducted based on the AQI values. Tables 1 and 2 in the study delineate the AQI index with corresponding pollutant concentration limits and the AQI classification, respectively. These tables provide a framework for understanding the correlation between pollutant concentrations and air quality categories, facilitating a comprehensive analysis of air quality. This analysis considers various factors, including pollutant concentration, exposure duration, and associated health risks, to assess air quality and categorize it appropriately.

Table 1. AQI index and pollutant concentration limits

IAQI	Concentration Limits for Pollutants									
	SO_2 24-Hour Average (μg/m^3)	SO_2 1-Hour Average (μg/m^3) （1）	NO_2 24-Hour Average (μg/m^3)	NO_2 1-Hour Average （μg/m^3） （1）	PM_{10} 24-Hour Average (μg/m^3)	CO 24-Hour Average (mg/m^3)	CO 1-Hour Average (mg/m^3) （1）	O_3 1-Hour Average (μg/m^3)	O_3 8-Hour Sliding Average (μg/m^3)	$PM_{2.5}$ 24-Hour Average (μg/m^3)
0	0	0	0	0	0	0	0	0	0	0
50	50	150	40	100	50	2	5	160	100	35
100	150	500	80	200	150	4	10	200	160	75
150	475	650	180	700	250	14	35	300	215	115
200	800	800	280	1200	350	24	60	400	265	150
300	1600	(2)	565	2340	420	36	90	800	800	250
400	2100	(2)	750	3090	500	48	120	1000	(3)	350
500	2620	(2)	940	3840	600	60	150	1200	(3)	500

Table 2. AQI classification

AQI	AQI Level	AQI Categories and Colors
0-50	1	Optimal (green)
51-100	2	Good (yellow)
101-150	3	Slight pollution (orange)
151-200	4	Moderate pollution (red)
201-300	5	Heavy pollution (purple)
>300	6	Severe pollution (maroon)

The evaluation of air quality levels and categories is systematically conducted based on the AQI values. The AQI serves as a quantitative tool, facilitating the evaluation of ambient air quality through a comprehensive analysis. This analysis incorporates several critical factors, including the concentration of various pollutants in the air, the duration of exposure, and the potential health risks associated with such exposure. The AQI values are then utilized to categorize the air quality into distinct levels, ranging from "optimal" to "severe pollution." It is imperative for businesses and academic institutions to monitor air quality levels diligently, as poor air quality poses significant risks to both human health and the environment.

3.2 BP Neural Network

3.2.1 Neuron model

In the BP Neural Network, each neuron functions by assigning a specific weight value to each incoming signal, determining the activation of the neuron. The collective weighted sum of these input signals is computed, ascertaining the neuron's activation status. These weights are indicative of the synaptic connection strength, a critical aspect of neural processing. Figure 1 illustrates the basic neuron model within this context.

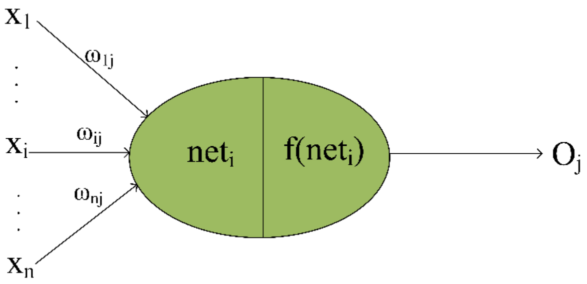

Figure 1. Neuron model

Within the neural network, each neuron, serving as a computational element, assimilates input signals, each represented by distinct connection weights. This neuron conducts a comprehensive evaluation of the input signals, culminating in the network's output. The aggregation of these input values is executed for each neuron, collectively contributing to the overall network output.

3.2.2 Basic principle of BP Neural Network

The architecture of the BP Neural Network comprises an input layer, one or more hidden layers, and an output layer. Initially, the network calculates the error through a process of forward propagation. Following this, the weights and thresholds of both the hidden and output layers are adjusted via backpropagation. This adjustment aims to minimize the mean square error. Figure 2 displays the topology of a BP Neural Network, specifically highlighting a configuration with two hidden layers.

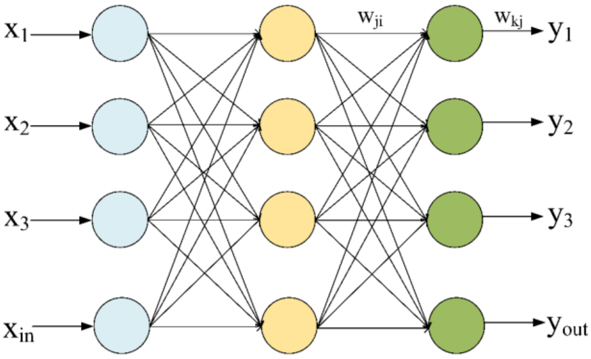

Figure 2. Topology structure of BP Neural Network

3.2.3 Propagation process of BP Neural Network

Learning in a neural network encompasses two principal stages: forward propagation and error backpropagation. During forward propagation, the network assimilates the weight and threshold values. Should the network's output deviate from the expected result, it transitions to error backpropagation. In this phase, an error function gradient descent strategy is employed to adjust the network's weights and thresholds, based on the magnitude of the relative error. The backpropagation process proceeds in the reverse direction of forward propagation, continuing until the BP Neural Network's output data closely aligns with the anticipated output. Figure 3 shows the topology of a three-layer neural network.

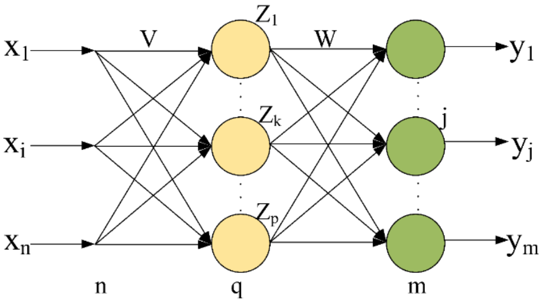

Figure 3. Topology of a three-layer neural network

(a) Forward propagation process of the signal

Consider a three-layer BP network comprising n input nodes, m output nodes, and a hidden layer with q nodes. The weights between the input and hidden layers are denoted as v_{ik}, and those between the hidden and output layers as w_{kj}. The transfer functions for the hidden and output layers are f_1 and f_2, respectively. The output equations for these layers are expressed as follows:

The output of the hidden layer node is calculated using Eq. (3):

$$Z_k = f_1\left(\sum_{i=0}^{n} V_{ik} X_i\right) k = 1, 2, \ldots q \tag{3}$$

The output of the output layer node is derived using Eq. (4):

$$y_j = f_2\left(\sum_{k=0}^{q} w_{kj} z_k\right) j = 1, 2, \ldots m \tag{4}$$

(b) Backpropagation process of error

Step 1: Error function definition

For a set comprising P samples, $x_1, x_2, x_3, \dots x_p$ is defined as the input data. Value of the output node $\mathrm{y}_\mathrm{j}^\mathrm{p}$ $(j = 1,2, \dots m)$ is obtained following the processing of the sample p by the neural network. The squared error function E_P for sample x_p is defined in Eq. (5):

$$\mathrm{E_P} = \frac{1}{2}\sum_{\mathrm{j=1}}^{\mathrm{m}}\left(\mathrm{t_j^p} - \mathrm{y_j^p}\right)^2 \tag{5}$$

where, t_j^p is the desired output. The global error for this sample is then calculated using Eq. (6):

$$\mathrm{E} = \frac{1}{2}\sum_{\mathrm{P=1}}^{\mathrm{P}}\sum_{\mathrm{j=1}}^{\mathrm{m}}\left(\mathrm{t_j^p} - \mathrm{y_j^p}\right)^2 = \sum_{\mathrm{P=1}}^{\mathrm{P}}\mathrm{E_P} \tag{6}$$

Step 2: Weight adjustment of output layer

The cumulative error is adjusted using the BP algorithm as per Eq. (7), with η representing the learning rate.

$$\Delta \mathrm{w_{jk}} = -\eta\frac{\partial \mathrm{E}}{\partial \mathrm{w_{jk}}} = -\eta\frac{\partial}{\partial \mathrm{w_{jk}}}\left(\sum_{\mathrm{p=1}}^{\mathrm{p}}\mathrm{E_P}\right) = \sum_{\mathrm{p=1}}^{\mathrm{p}}\left(-\eta\frac{\partial \mathrm{E_P}}{\partial \mathrm{w_{jk}}}\right) \tag{7}$$

The error signal is defined in Eq. (8), further detailed in items in Eqs. (9) and (10):

$$\delta_{yj} = -\frac{\partial \mathrm{E_P}}{\partial \mathrm{S_j}} = -\frac{\partial \mathrm{E_P}}{\partial \mathrm{y_j}}\cdot\frac{\partial \mathrm{y_j}}{\partial \mathrm{S_j}} \tag{8}$$

$$\frac{\partial \mathrm{E_p}}{\partial \mathrm{y_j}} = \frac{\partial}{\partial \mathrm{y_j}}\left[\frac{1}{2}\sum_{\mathrm{j=1}}^{\mathrm{m}}\left(\mathrm{t_j^p} - \mathrm{y_j^p}\right)^2\right] = -\sum_{\mathrm{j=1}}^{\mathrm{m}}\left(\mathrm{t_j^p} - \mathrm{y_j^p}\right) \tag{9}$$

$$\frac{\partial y_j}{\partial s_j} = f_2'\left(S_j\right) \tag{10}$$

The partial differentiation of the transfer function for the output layer is given in Eq. (11):

$$\delta_{yj} = \sum_{j=1}^{m}\left(t_j^p - y_j^p\right)f_2'\left(S_j\right) \tag{11}$$

Applying the chain theorem, Eq. (12) is derived, leading to the formulation of Eq. (13) for adjusting the weights of neurons in the output layer.

$$\frac{\partial \mathrm{E_P}}{\partial \mathrm{w_{jk}}} = \frac{\partial \mathrm{E_P}}{\partial \mathrm{S_j}}\cdot\frac{\partial \mathrm{S_j}}{\partial \mathrm{w_{jk}}} = -\delta_{\mathrm{yj}}\mathrm{Z_k} = -\sum_{\mathrm{j=1}}^{\mathrm{m}}\left(\mathrm{t_j^p} - \mathrm{y_j^p}\right)\mathrm{f_2'}\left(\mathrm{S_j}\right)\mathrm{Z_k} \tag{12}$$

$$\Delta w_{jk} = \sum_{p=1}^{p}\sum_{j=1}^{m}\eta\left(t_j^p - y_j^p\right)f_2'\left(S_j\right)Z_k \tag{13}$$

The weight adjustment for the hidden layer is executed using Eq. (14):

$$\Delta \mathrm{v_{ki}} = -\eta\frac{\partial \mathrm{E}}{\partial \mathrm{v_{ki}}} = -\eta\frac{\partial}{\partial \mathrm{v_{ki}}}\left(\sum_{\mathrm{p=1}}^{\mathrm{p}}\mathrm{E_p}\right) = \sum_{\mathrm{p=1}}^{\mathrm{p}}\left(-\eta\frac{\partial \mathrm{E_p}}{\partial \mathrm{v_{ki}}}\right) \tag{14}$$

Repeating the aforementioned steps, Eq. (15) is obtained, facilitating the comprehensive weight adjustment process for the hidden layer.

$$\Delta v_{ki} = \sum_{p=1}^{p}\sum_{j=1}^{m}\eta\left(t_j^p - y_j^p\right) f_2'\left(S_j\right) w_{jk} f_1'\left(S_k\right) x_i \tag{15}$$

3.3 K-means Cluster Analysis

In this study, K-means cluster analysis was employed to categorize research objects based on their attributes. The essence of K-means lies in iteratively finding a partition scheme of K clusters that minimizes the associated cost function, which is defined as the sum of squared errors of the distances between each sample and its corresponding cluster center. Due to its scalability and near-linear computational complexity, K-means is particularly adept at handling large datasets. While the algorithm may converge to local optima, these are generally sufficient for clustering objectives. The application of K-means in this research was augmented with geographic information to delineate the spatial distribution of haze clusters.

3.4 Data Sources

The data for this research was compiled from various authoritative sources within China, encompassing the Atmospheric Administration, National Aeronautics and Space Administration (NASA), and the Tianjin Bureau of Environmental Statistics. The dataset spans from January 1, 2014, to December 31, 2017, and predominantly includes daily air pollution data from 13 cities in China. These cities are Baoding, Beijing, Cangzhou, Chengde, Handan, Hengshui, Langfang, Qinhuangdao, Shijiazhuang, Tangshan, Tianjin, Xingtai, and Zhangjiakou. The collected data comprises AQI, air quality levels, and concentrations of various pollutants including $PM_{2.5}$, PM_{10}, sulfur dioxide, nitrogen dioxide, carbon monoxide, and ozone. Table 3 provides a snapshot of the air quality data for Beijing in 2017, illustrating the AQI, air quality levels, and concentrations of $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3 on select dates.

Table 3. Partial air quality data for Beijing in 2017

Date	AQI	Air Quality Level	$PM_{2.5}$	PM_{10}	SO_2	NO_2	CO	O_3
2017/1/1	454	Severe pollution	430	501	8	131	6.43	4
2017/2/1	54	Good	30	49	12	24	0.6	56
2017/3/1	44	Optimal	7	37	2	13	0.52	79
2017/4/1	61	Good	28	73	5	50	0.58	48
2017/5/1	82	Good	40	112	7	44	0.7	71
2017/6/1	69	Good	22	56	3	27	0.6	113
2017/7/1	137	Slight pollution	89	132	3	40	0.97	154
2017/8/1	98	Good	67	86	2	29	0.96	92
2017/9/1	177	Heavy pollution	134	149	2	38	1.39	71
2017/10/1	111	Slight pollution	73	91	1	47	0.75	52
2017/11/1	88	Good	63	99	2	76	1.01	12
2017/12/1	91	Good	67	99	10	66	1.26	12

4. Result Analysis

4.1 AQI of Cities in the Beijing-Tianjin-Hebei Region

The evaluation of air quality in the Beijing-Tianjin-Hebei region was conducted using AQI. An AQI value of 100 or below is indicative of 'excellent' or 'good' air quality. Figure 4 illustrates a statistical chart that presents the frequency of days with good air quality and the number of polluted days in this region.

Analysis of Figure 4 reveals that certain cities, including Zhangjiakou, Chengde, and Qinhuangdao, have maintained a 'good' air quality rate exceeding 80%. Conversely, other cities in the region exhibited lower 'good' air quality rates, falling below the 80% threshold. Notably, cities such as Shijiazhuang, Handan, Baoding, Hengshui, and Xingtai reported 'good' air quality rates below 60%. This data suggests that these cities endure substantial air pollution for nearly half of the year.

The significance of air pollution on environmental and human health, especially in regions with suboptimal air quality, cannot be overstated. It is, therefore, vital to develop and implement effective strategies to mitigate air pollution and protect public health.

Figure 5 provides an insightful visual representation of air quality levels across the Beijing-Tianjin-Hebei region for the year 2017. This graphical depiction offers a detailed overview of the regional air quality, facilitating the assessment of air quality status and the development of targeted air quality improvement policies.

According to the air quality classification diagram, Chengde and Zhangjiakou cities exhibit the highest percentage of days classified as 'excellent' air quality, with 61% of days achieving 'good' levels. In contrast, cities such as Baoding, Handan, Hengshui, Shijiazhuang, and Xingtai experience over 25% of days with light pollution, and a significant proportion of days exhibit moderate to heavy pollution. This pattern indicates a more severe pollution situation in these cities. It is evident that the 13 cities within the Beijing-Tianjin-Hebei region confront diverse pollution challenges that are persistent both temporally and spatially.

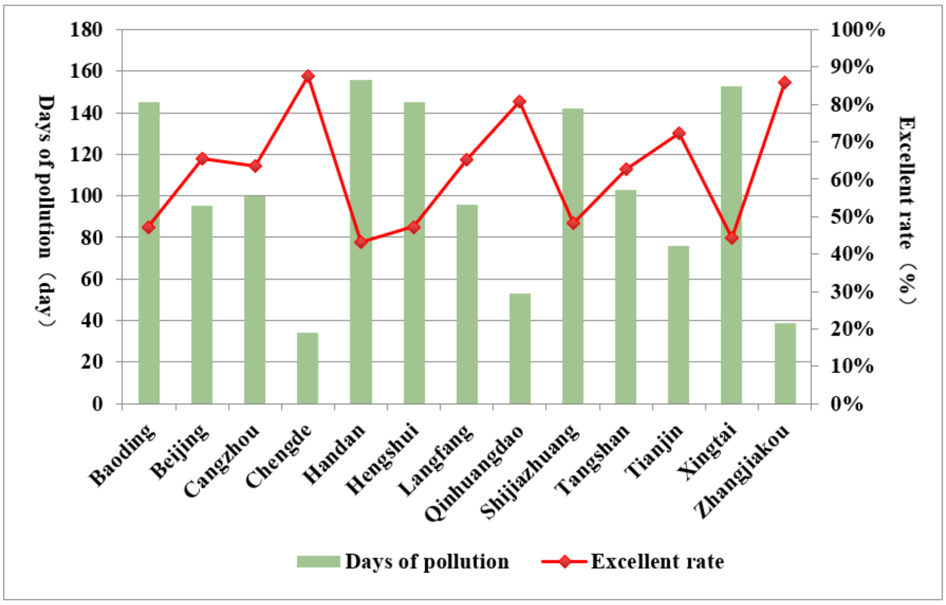

Figure 4. Statistical chart of good air quality rates and pollution days in the Beijing-Tianjin-Hebei region

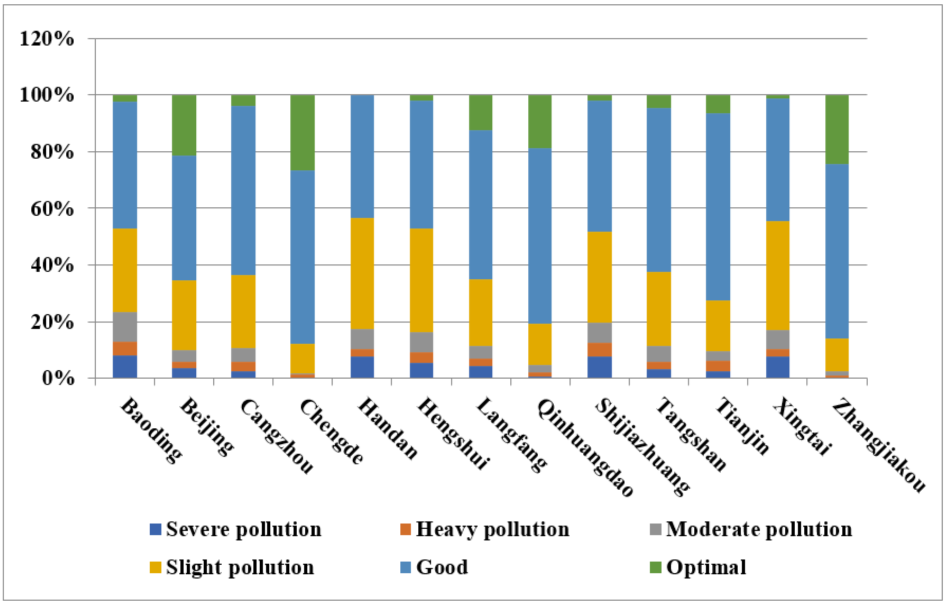

Figure 5. Air quality class distribution in the Beijing-Tianjin-Hebei region

4.2 Monthly Variation Rules of Air Quality

In the Beijing-Tianjin-Hebei region, the highest AQI values are typically recorded from January to March and from October to December, with the peak period occurring in January, February, November, and December (Figure 6). During the summer months, characterized by low atmospheric pressure, the region experiences its highest temperatures, increased humidity, and more dynamic atmospheric conditions, which enhance the likelihood of precipitation and wind activity. In contrast, the winter season, governed by high atmospheric pressure, witnesses the lowest temperatures and more stable weather patterns, reducing the chances of precipitation and wind. Meteorological conditions have been identified as the primary influencers of the temporal distribution of smog in this region.

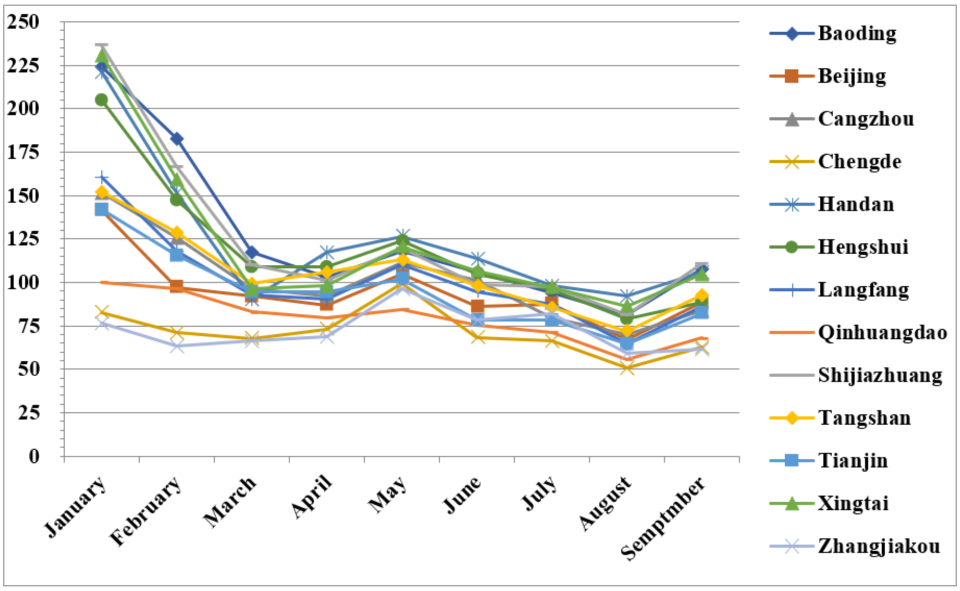

Figure 6. Trend of monthly mean AQI concentration in 2017

4.3 AQI Spatial Distribution of Seasonal Variations

The spatial distribution of haze pollution across the 13 cities in the Beijing-Tianjin-Hebei region was investigated using quarterly average AQI data from 2017. GeoDa software was employed to generate cluster distribution maps of AQI for each season, as depicted in Figure 7.

The diagrams in Figure 7 indicate that an AQI value of 1 represents high pollution levels, whereas a value of 5 signifies low pollution. Comparative analysis of AQI values across the four seasons reveals distinct seasonal variations in haze pollution. Summer is characterized by relatively good air quality, with an escalation in haze coverage commencing in autumn and peaking in winter. The clustering results suggest that cities in close proximity exhibit similar pollution levels, highlighting the presence of local spatial clustering in haze distribution. This finding underscores the necessity for collaborative efforts among neighboring cities to effectively address air pollution.

Moreover, the pollution gradient tends to increase from northeast to southwest, with notable disparities between the northern and southern regions. Cities in the north consistently face heavy pollution throughout the year, posing significant health risks from prolonged exposure to polluted air. The southern cities, while experiencing heightened pollution levels in spring and autumn, witness exacerbated conditions in winter. This seasonal variation in pollution, especially the heightened levels in northern cities like Qinhuangdao, Chengde, and Langfang during winter, significantly influences the clustering patterns. Consequently, understanding the underlying causes of these seasonal changes is crucial for formulating effective air quality improvement strategies.

4.4 Prediction of AQI

Following the evaluation of the trained neural network, the results demonstrated an average accuracy rate of 83.94% and an average mean squared error (MSE) of 0.0247. These results are detailed in Table 4.

The findings are elucidated in Figure 8, which graphically represents the training results of the partial grid for the prediction model.

A BP Neural Network model was developed to predict and analyze the haze data from 2017. The model's predicted AQI values were juxtaposed with the actual recorded values, as exemplified by the AQI prediction results for Beijing in 2017 shown in Figure 9. This comparison is critical in evaluating the model's effectiveness in forecasting AQI values.

Inspection of Figure 9 reveals a close correlation between the predicted and actual AQI values. The congruence in the changing trend underscores the efficacy of the model. The primary objective of this analysis is to observe and provide alerts for sudden changes in AQI values. As indicated in Figure 9, the model successfully identified significant AQI increases, thereby proving its utility in offering predictive alerts for significant changes in air quality.

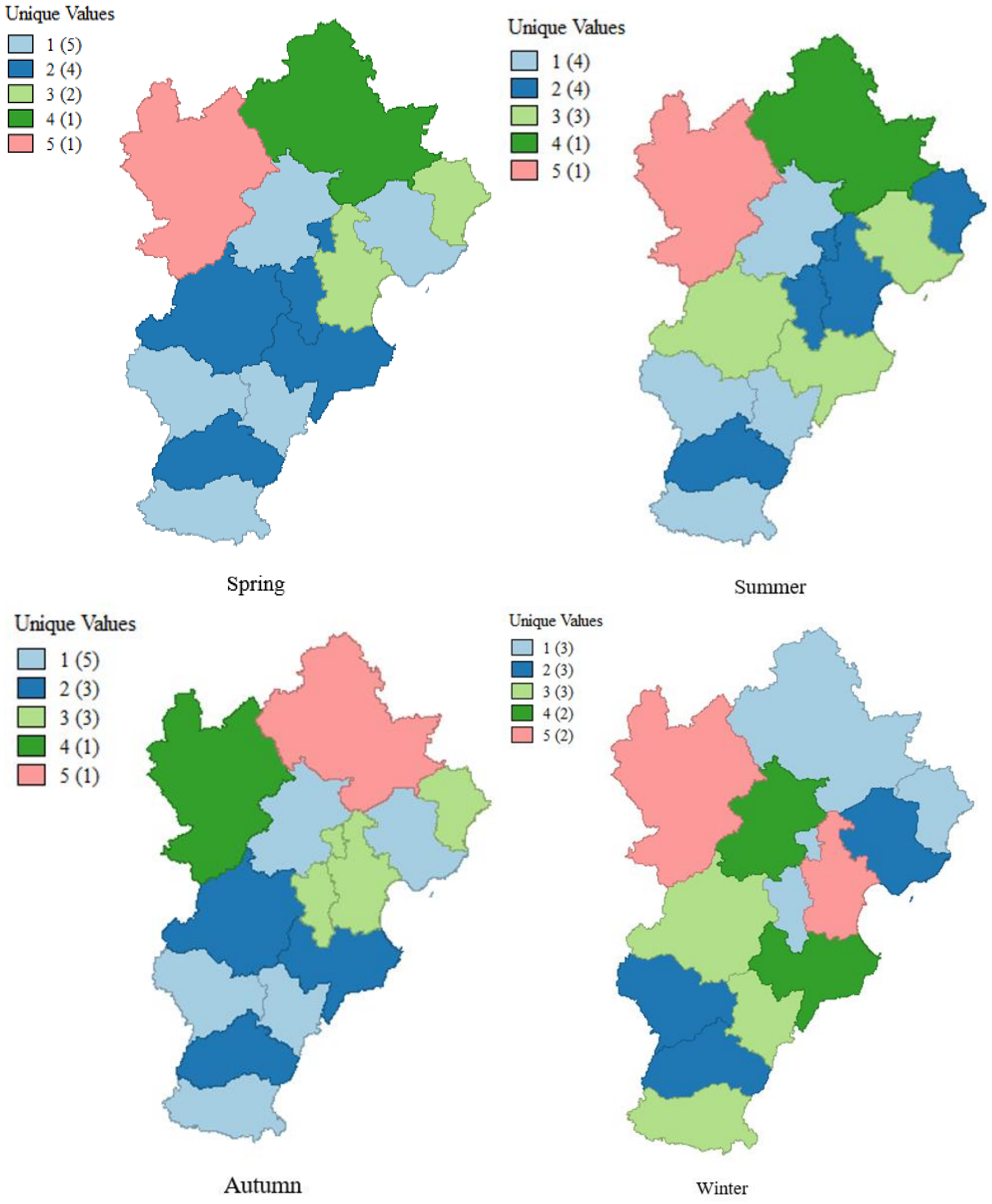

Figure 7. Seasonal distribution of AQI in the Beijing-Tianjin-Hebei region in 2017

Table 4. Results on a partial test set

Test Set	MSE	Accuracy
1	0.0325	82.6%
2	0.0213	86.0%
3	0.0218	85.6%
4	0.0223	85.5%
5	0.0285	82.8%
6	0.0218	85.6%
7	0.0202	86.2%
…	…	…
3796	0.0204	86.1%
Average	0.0247	83.94%

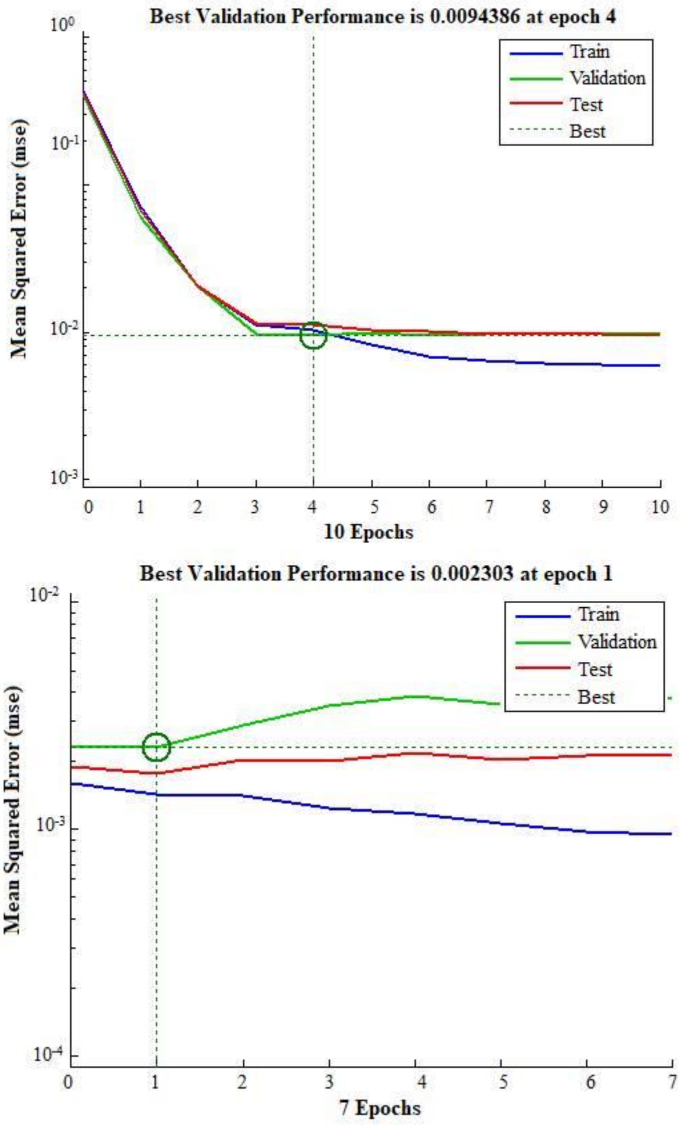

Figure 8. BP Neural Network training results

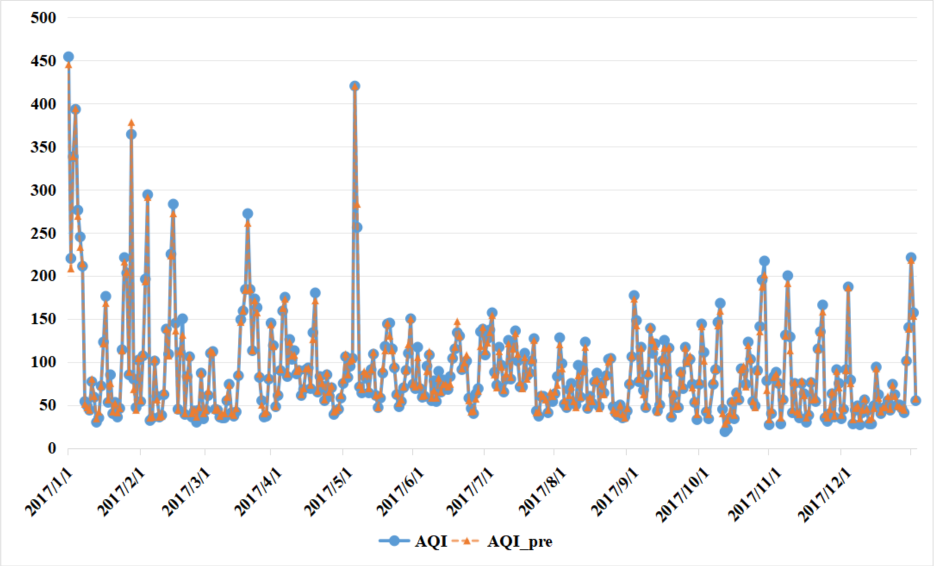

Figure 9. Comparison of predicted and actual AQI values in Beijing for 2017

5. Conclusion

The challenge of mitigating haze pollution holds significant implications for China's environmental sustainability and economic progress. This study, focusing on 13 prefecture-level cities within the Beijing-Tianjin-Hebei region, provides a comprehensive analysis of air quality data to elucidate the spatial distribution patterns of haze pollution in China. The following conclusions are drawn:

First, AQI serves as an effective direct measure of air quality. In this study, six principal influencing factors were selected as inputs for the BP Neural Network to forecast AQI values.

Second, analysis using the K-means clustering method reveals that haze pollution in the Beijing-Tianjin-Hebei region exhibits local spatial aggregation with distinct seasonal variations. The study finds that air quality is most favorable in summer, with minimal pollution levels, while winter experiences heightened haze pollution across a broader area. Notably, the southern part of Hebei Province is identified as the most polluted area. These distribution patterns are intricately linked to the regional topography and socio-economic factors.

Third, the BP Neural Network model's prediction of AQI demonstrates satisfactory accuracy, particularly in instances of significant AQI fluctuations. The model's predictive capabilities align closely with actual trends, offering valuable insights for issuing pollution alerts and reminders based on forecasted AQI values.

In summation, this study explores the temporal distribution of AQI, analyzing daily, monthly, and quarterly variations. It employs K-means clustering analysis to study the seasonal spatial distribution of haze pollution, utilizing AQI as a multifaceted evaluation metric. Moreover, the study harnesses the predictive power of the BP Neural Network for AQI forecasting, providing critical information for public alerts on air quality fluctuations to facilitate safer travel and living conditions.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

Abdul-Rahman, T., Roy, P., Bliss, Z. S. B., Mohammad, A., Corriero, A. C., Patel, N. T., Wireko, A.A., Shaikh, R., Faith, O.E., Arevalo-Rios, E.C., Dupuis, L., Ulusan, S., Erbay, M.I., Cedeno, M.V., Sood, A., & Gupta, R. (2024). The Impact of Air Quality on Cardiovascular Health: A state of the art review. *Curr. Probl. Cardiol.*, *49*(2), 102174. https://doi.org/10.1016/j.cpcardiol.2023.102174.

Ahmad, S. & Ahmad, T. (2023). AQI prediction using layer recurrent neural network model: A new approach. *Environ. Monit. Assess*., *195*(10), 1180. https://doi.org/10.1007/s10661-023-11646-3.

Cai, X., Hu, H., Liu, C., Tan, Z., Zheng, S., & Qiu, S. (2023). The effect of natural and socioeconomic factors on haze pollution from global and local perspectives in China. *Environ Sci. Pollut R.*, *30*(26), 68356-68372. https://doi.org/10.1007/s11356-023-27134-7.

Duangsuwan, S., Prapruetdee, P., Subongkod, M., & Klubsuwan, K. (2022). 3D AQI mapping data assessment of low-altitude drone real-time air pollution monitoring. *Drones*, *6*(8), 191. https://doi.org/10.3390/drones6080191.

Jia, P. & Yan, J. (2022). Effects of haze pollution and institutional environment on demand for commercial health insurance. *Front. Psychol.*, *13*, 1002470. https://doi.org/10.3389/fpsyg.2022.1002470.

Li, J. (2018). Causes of formation of haze and its control measures based on cluster analysis. *IPPTA: Quart. J. Ind. Pulp Pap. Tech. Assoc.*, *30*(8), 263-267.

Li, Z., Zhou, J., & Zhang, Z. (2023). Market segmentation and haze pollution in Yangtze River Delta urban agglomeration of China. *Atmos.*, *14*(10), 1539. https://doi.org/10.3390/atmos14101539.

Liu, X. & Guo, H. (2022). Air quality indicators and AQI prediction coupling long-short term memory (LSTM) and sparrow search algorithm (SSA): A case study of Shanghai. *Atmos. Pollut. Res.*, *13*(10), 101551. https://doi.org/10.1016/j.apr.2022.101551.

Liu, Y. Z., Ren, T. T., Liu, L. J., Ni, J. L., & Yin, Y. K. (2023a). Heterogeneous industrial agglomeration, technological innovation and haze pollution. *China Econ. Rev.*, *77*, 101880. https://doi.org/10.1016/j.chieco.2022.101880.

Liu, Z., Lin, J., Zhou, L., Song, Y., & Li, X. (2023b). Applied research on AQI prediction based on BP Neural Network modeling. *Acad. J. Comput. Inf. Sci.*, *6*(10), 93-99. https://doi.org/10.25236/AJCIS.2023.061014.

Lv, Y. Q., Fan, T. Z., Zhao, B., Zhang, J., Zheng, Y., & Zhang, Z. (2022). How do government environmental concerns affect haze pollution? *Front. Environ. Sci.*, *10*, 945226. https://doi.org/10.3389/fenvs.2022.945226.

Ma, T. & Cao, X. (2022). The effect of the industrial structure and haze pollution: Spatial evidence for China. *Environ Sci. Pollut R.*, *29*(16), 23578-23594. https://doi.org/10.1007/s11356-021-17477-4.

Si, Y., Chen, L., Zheng, Z., Yang, L., Wang, F., Xu, N., & Zhang, X. (2023). A novel algorithm of haze identification based on FY3D/MERSI-II remote sensing data. *Remote Sens.*, *15*(2), 438. https://doi.org/10.3390/rs15020438.

Supphapipat, K., Leurcharusmee, P., Chattipakorn, N., & Chattipakorn, S. C. (2023). Impact of air pollution on postoperative outcomes following organ transplantation: Evidence from clinical investigations. *Clin. Transplant.*, e15180. https://doi.org/10.1111/ctr.15180.

Susilo, Y. S. & Putranto, L. F. D. (2023). Several variables affecting provincial Air Quality Index (AQI) in Indonesia 2012–2019. *in IOP Conference Series: Earth and Environmental Science, Volume 1180, International Conference on Environmental Management 2022 23/09/2022 - 24/09/2022 Online, Indonesia*, *1180*, 012041. https://doi.org/10.1088/1755-1315/1180/1/012041.

Van, D. A., Vu, T. V., Nguyen, T. H. T., Vo, L. H. T., Le, N. H., Nguyen, P. H., Pongkiatkul, P., & Ly, B. T. (2022). A review of characteristics, causes, and formation mechanisms of haze in Southeast Asia. *Curr. Pollut. Rep.*, *8*(2), 201-220. https://doi.org/10.1007/s40726-022-00220-z.

Wang, J. & Xu, Y. B. (2022). How does digitalization affect haze pollution? The mediating role of energy consumption. *Int. J. Environ. Res. Public Health.*, *19*(18), 11204. https://doi.org/10.3390/ijerph191811204.

Wu, W. (2023). Is air pollution joint prevention and control effective in China-evidence from "Air Pollution Prevention and Control Action Plan". *Environ Sci. Pollut R.*, 1-15. https://doi.org/10.1007/s11356-023-30982-y.

Xu, Y. Z., Zhang, R. J., Dong, B. Y., & Wang, J. (2023). Can the construction of low-carbon cities reduce haze pollution? *J. Environ. Plan. Manag.*, *66*(3), 590-620. https://doi.org/10.1080/09640568.2021.2000372.

Zhang, Q., Tian, L.F., Fu, F.C., Wu, H.Y., Wei, W., & Liu, X.Y. (2022). Real-time and image-based AQI estimation based on deep learning. *Adv. Theory Simul.*, *5*(6), 2100628. https://doi.org/10.1002/adts.202100628.