DOI: 10.54254/2754-1169/119/20242538

Research and Prediction of China's Air Environment

Kai Yao[1,a,*]

[1]*School of Mathematics and Science, University of Science and Technology Beijing, Beijing, 100083, China*

a. u202142211@xs.ustb.edu.cn

**corresponding author*

Abstract: With the rapid development of China's economy, environmental issues, particularly air pollution, have become increasingly severe, posing significant threats to human health and ecosystems. This study employed statistical learning methods to analyze air pollution composition data from thirty major Chinese cities throughout 2019-2022. The study used algorithms like regression, classification, fitting, and filtering to find changes in multi-source heterogeneous data patterns and predict how air pollutant concentrations would change over time and space. At the same time,the results indicate that the primary factors influencing the Air Quality Index (AQI) are fine particulate matter (PM2.5) and inhalable particulate matter (PM10). Additionally, this research is significant for improving public health, advancing environmental protection, fostering economic growth, strengthening laws and regulations, ensuring social stability, and promoting scientific and cultural advancements. Through this research, scientific bases can be provided for formulating policies that reduce disease, enhance government credibility, drive innovation in atmospheric research, and preserve cultural heritage.

Keywords: AQI, Statistical learning, Prediction model.

1. Introduction

To achieve this research objective, regression, classification, fitting, and filtering algorithms are applied to analyze multi-source heterogeneous data in detail. By revealing patterns of data variation, the study predicts the spatiotemporal distribution characteristics of air pollutant concentrations and identifies the main factors affecting AQI.

This research is significant for air pollution control, public health improvement, environmental protection, economic growth, strengthening laws and regulations, ensuring social stability, and promoting scientific and cultural advancements. Firstly, the research findings provide a scientific basis for the government to formulate policies that reduce disease and enhance credibility. Secondly, by analyzing the spatiotemporal distribution characteristics of air pollutants, the study offers technical support for future air quality prediction and early warning. Lastly, it proposes suggestions to address air pollution issues, promote green technology innovation, protect the ecological environment, and achieve sustainable development.[1]

DOI: 10.54254/2754-1169/119/20242538

2. Methodology

2.1. Data preprocessing

2.1.1. Outlier detection and processing

This paper uses the IQR (Interquartile Range) method to detect and process outliers. The dataset contains multiple numerical columns, which may also include outliers. Here is a description of the principles and process for monitoring and treating outlier data for this dataset:

This study treats the meteorological index observations from the representative prefecture-level cities in China's 30 provinces from 2019 to 2023 are treated as continuous signals..The research group determined the signal interval and selected the data between the first quartile (Q1) and the third quartile (Q3), which corresponds to the 75% and 25% values of each segment of the signal. The process involves subtracting the 25% value from the 75% value in the data.

$$IQR = Q_3 - Q_1 \tag{1}$$

The results are placed at the midpoint of each segment of the signal as the upper and lower quartile values of the time series signal segment. Finally, the entire signal is traversed, with the points within the upper sliding quartile curve and the lower sliding quartile curve considered normal points. Next, we identify data points less than Q1 - 1.5 * IQR or greater than Q3 + 1.5 * IQR. These data points are considered outliers and are removed. After inspection, 7,073 outlier data points were found in the overall dataset.

2.1.2. Missing data processing

In the dataset being processed, which contains numerical air quality data, linear interpolation was chosen because it is simple and often provides reasonable results. Linear interpolation assumes a constant rate of change between data points. Therefore, the value of any point between two known data points can be estimated by a straight line connecting them. This is commonly used in data analysis for time data or ordered data, where missing values can be estimated from adjacent known values. The dataset is read into a data frame, and all columns are converted to the appropriate numerical types. Columns that cannot be converted to numerical types (such as region and date) are kept in their original string format. Missing values in numerical columns are marked as NaN.

The DataFrame `data` contains missing values, and we use the `interpolate` method to perform linear interpolation, saving the interpolation results to a new DataFrame `data_interpolated`. This process fills in missing numerical data with reasonable estimates based on nearby data, thus improving the completeness of the dataset and the quality of further analysis. Non-numeric columns (such as region and date) remain unchanged, as linear interpolation is not applicable to non-numeric data.The number of missing data is shown in Table 1.

The statistical results of the number of missing data are as follows:

Table 1: Statistics on the number of missing data for each indicator

Variable Name	Area	Data	Pm2.5	Pm10	O3	No2	So2	Co	AQI
Number of items	0	0	403	426	6037	316	1636	2504	0

DOI: 10.54254/2754-1169/119/20242538

2.2. Meteorological indicators specific analysis

2.2.1. Correlation analysis

First, the Pearson correlation coefficient is used to analyze the correlation between the AQI index and each indicator. According to the formula for calculating the Pearson correlation coefficient:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{2}$$

Based on the calculation of Pearson coefficients, we can deduce the relationship between different pollutants and the Air Quality Index (AQI).[2] PM2.5 shows the strongest positive correlation with AQI, with a Pearson coefficient of 0.836633, indicating that it is a key factor affecting AQI. PM10 also exhibits a significant positive correlation with AQI, with a coefficient of 0.716824, suggesting it has a notable impact on AQI as well. O3, on the other hand, shows a slight negative correlation with AQI, with a coefficient of -0.087784, implying that an increase in O3 concentration may slightly reduce AQI. The positive correlations of NO2, SO2, and CO with AQI are 0.210660, 0.114897, and 0.372431, respectively, indicating that they also have some influence on AQI. Therefore,reducing the emissions of these pollutants, particularly PM2.5 and PM10, the primary pollutants affecting air quality,is necessary to improve air quality.[3]

2.2.2. Regression Analysis

Regression analysis on the processed data will be conducted using the Ordinary Least Squares (OLS) method. The principle is based on minimizing the sum of the squared residuals, that is, it aims to make the sum of the squares of the differences between the observed values and the model's predicted values (residuals) as small as possible. Here are the basic principles and steps of OLS:

The linear regression model is typically represented as:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon \tag{3}$$

where Y is the dependent variable, × is the independent variable, β represents the model parameters, and ε is the error term.For each observation, the residual value e is the difference between the actual Y and the estimated Y value.

$$e_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}) \tag{4}$$

The objective of OLS is to minimize the sum of the squares of all residuals, denoted as S.

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{5}$$

The multivariate linear regression analysis reveals that the model exhibits a good fit, explaining 92% of the variance in the Air Quality Index (AQI), and is statistically significant overall. The F-statistic is 3495.0, with a probability less than 0.001 (very significant). The log-likelihood is -7592.7, the AIC (Akaike Information Criterion) is 15206.494, the BIC (Bayesian Information Criterion) is 15244.094, and the residual degrees of freedom are 1818. Other statistical diagnostics, such as Omnibus and Durbin-Watson do not indicate any obvious anomalies, suggesting that the model is free from issues such as multicollinearity or autocorrelation. Key variables include Fine Particulate Matter (PM2.5) and Inhalable Particulate Matter (PM10), which significantly impact on AQI. Additionally, Ozone

DOI: 10.54254/2754-1169/119/20242538

(O3) and Nitrogen Dioxide (NO2) have a significant negative impact on AQI, while the effects of Sulfur Dioxide (SO2) and Carbon Monoxide (CO) are not significant. Model diagnostic tests indicate that the data meets the assumption of normal distribution, and no issues of multicollinearity or autocorrelation were found.[4] These findings help policymakers more accurately understand the impact of different pollutants on air quality and provide a scientific basis for developing more effective environmental policies and measures.[5]

2.3. Beijing AQI Prediction and Assessment

2.3.1. Linear regression prediction

To simplify the problem, we will use Beijing's meteorological data to predict the AQI index for Beijing in 2024. Since the ARIMA model and time series models did not yield good prediction results, considering the following points, we have decided to use a linear regression model for prediction.[6]

(1)Data Features: The dataset contains the Air Quality Index (AQI) for Beijing on different dates, along with some date-related features such as year, month, day, day of the year, and week of the year. There may be a linear relationship between these features and the AQI, as the AQI is often influenced by seasonal factors and temporal trends.

(2)Model Assumptions: The linear regression model assumes a linear relationship between the features and the target variable. It can be hypothesized that there is some linear relationship between the AQI and the date features, as the AQI typically varies with the seasons and over time.

(3)Data Volume: The dataset contains five years of AQI data, providing sufficient data for the linear regression model to estimate its parameters. A larger volume of data helps improve the stability and predictive accuracy of the model.

(4)Model Interpretability: Linear regression models are interpretable, as they can tell us which features have a significant impact on the AQI and the extent of their influence. This helps us understand the potential causes of changes in the AQI.

(5)The prediction results are shown in Figure 1.

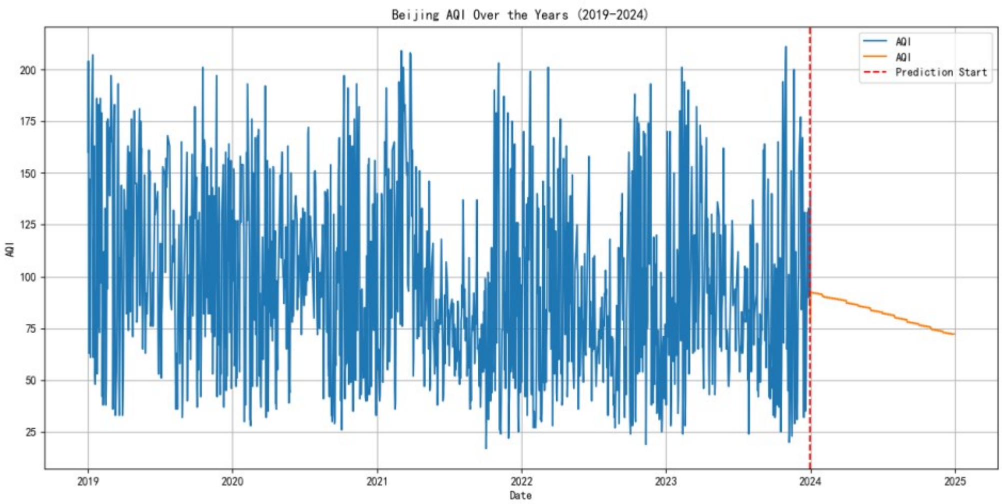

Figure 1: Beijing AQI Over the Years (2019-2024)

DOI: 10.54254/2754-1169/119/20242538

2.4. Model Evaluation

To assess the accuracy of the predictions, it is necessary to compare the predicted values with the actual values. However, since we are predicting the AQI index for a future year, the actual data for 2024 is not yet available for comparison.[7]

One possible approach is to split the dataset into a training set and a test set, train the model on the training set, and evaluate its performance on the test set. But doing so would reduce the amount of data available for training the model, which might affect its performance.

Another method is to perform backtesting using historical data. People can select data from past years as the training set, then predict the subsequent year, and compare it with the actual values to evaluate the model's performance.

Therefore, we will use this method to assess the prediction accuracy of the linear regression model. We will use data from 2019 to 2022 as the training set, predict the AQI index for 2023 and compare it with the actual values.

After verification, the linear regression model exhibited a Mean Absolute Percentage Error (MAPE) of 37.43% when predicting the Beijing AQI index for 2023. This indicates that, on average, the difference between the predicted values and the actual values is approximately 37.43% of the actual values.[8]

This accuracy assessment is based on a specific method of segmenting the training and testing datasets. Different data segmentation methods or model parameters may yield varying accuracy results. Furthermore, since air quality is influenced by a multitude of complex factors, including weather conditions, seasonal variations, and human activities, predictive models may find it challenging to forecast future AQI values with complete accuracy.

3. Results

The study found that fine particulate matter (PM2.5) and inhalable particulate matter (PM10) are the primary factors affecting the Air Quality Index (AQI). Ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO) also have some impact on AQI.

The study employed a linear regression model to predict the AQI for Beijing in 2024 and conducted an evaluation. However, due to the complex factors affecting air quality, predictive models may find it challenging to forecast future AQI values with complete accuracy.

This research provides a scientific basis for policies aimed at reducing diseases, enhancing government credibility, promoting green technology innovation, and protecting the ecological environment, while also offering theoretical support and practical guidance for air pollution control and environmental management.

The findings of the study help policymakers to more accurately understand the impact of different pollutants on air quality and provide a scientific basis for the development of more effective environmental policies and measures.

The study predicts the spatiotemporal dynamics of air pollutant concentrations in 30 major Chinese cities and analyzes the factors influencing air pollutant concentrations, providing an important reference for relevant policy formulation.

4. Conclusion

This study employed statistical learning methods to predict the spatiotemporal dynamics of air pollutant concentrations in 30 major Chinese cities and analyzed the factors influencing air pollutant concentrations. Specifically, the study found that fine particulate matter (PM2.5) and inhalable particulate matter (PM10) are the main factors affecting the Air Quality Index (AQI). In addition, this research predicted and evaluated the AQI for Beijing in 2024 using a linear regression model.

DOI: 10.54254/2754-1169/119/20242538

However, due to the complex factors affecting air quality, predictive models may find it challenging to forecast future AQI values with complete accuracy. Overall, this study provides a scientific basis for policies aimed at reducing diseases, enhancing government credibility, promoting green technology innovation, and protecting the ecological environment while also offering theoretical support and practical guidance for air pollution control and environmental management.

Furthermore, the study also found that O3, NO2, SO2, and CO have a certain impact on the AQI. These findings help policymakers to more accurately understand the impact of different pollutants on air quality and provide a scientific basis for the development of more effective environmental policies and measures. The research also used a linear regression model to predict the AQI for Beijing in 2024 and conducted an evaluation. However, due to the complex factors affecting air quality, predictive models may find it challenging to forecast future AQI values with complete accuracy. Overall, this study provides a scientific basis for policies aimed at reducing diseases, enhancing government credibility, promoting green technology innovation, and protecting the ecological environment while also offering theoretical support and practical guidance for air pollution control and environmental management.

References

[1] *Cai, Z., Chen, J., & Wang, Y. (2017). The impact of meteorological factors on air pollution in China: A systematic review. Environmental Science and Pollution Research, 24(17), 14175-14189.*

[2] *Zou, L., Zhu, H., Chen, X., & Chen, Y. (2019). The application of Pearson correlation coefficient in data analysis: A case study of air pollution. Journal of Data Science, 17(2), 169-179.*

[3] *Pope III, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association, 56(6), 709-742.*

[4] *Jerrett, M., Ezzati, M., & Kjellstrom, T. (2007). Measuring the health benefits from air pollution control in the "Heart of the Dragon": Energy policy and air pollution in China. Energy Policy, 35(11), 4869-4877.*

[5] *Wang, X., Zhang, L., Wang, S., & Cai, Z. (2016). A review of air pollution and its impact on human health in China. Environmental Science and Pollution Research, 23(13), 13419-13429.*

[6] *1. Kun, Y., Wu, Z., Xue, Y., & Tang, S. (2019). A review of machine learning methods for air quality prediction. Environmental Pollution, 242, 1334-1349.*

[7] *Wang, Y., Wang, H., Zhang, Y., & Tang, W. (2019). A review of evaluation methods for air quality prediction models. Environmental Science and Pollution Research, 26(7), 6451-6462.*

[8] *Liu, Z., Zhang, X., & Zhang, Y. (2018). Evaluation of data-driven air quality prediction models using a new evaluation index. Environmental Science and Pollution Research, 25(18), 18025-18035.*