

Study on Spatial and Temporal Characteristics Analysis and Prediction of Air Quality in Shanxi Province Under Multi-Model Fusion

Lenglu Wen

Tai yuan University, Taiyuan 030032, China

Abstract: This paper takes the air quality data of various cities in Shanxi Province from 2013 to 2024 as the research object and employs a multi-model integration analysis method for comprehensive evaluation and prediction of air quality. First, based on Air Quality Index (AQI) and major pollutant concentration data, SPSS software is used to complete KMO and Bartlett tests for data applicability analysis. After standardization processing, SPSS's factor analysis module is applied to extract key principal components and construct factor score equations. In terms of classification model construction, the study innovatively uses the Matlab time series toolbox to perform strict stationarity tests and parameter optimization, establishing an ARIMA(1,0,1) forecasting model. This model demonstrates excellent performance in both fit and prediction accuracy. The study shows that the multi-model integration analysis framework can effectively reveal the spatiotemporal characteristics of air quality and their influencing factors.

Keywords: Factor Analysis, ARIMA Prediction Model, KMO Test, Bartlett Test, Temporal and Spatial Characteristics.

1. Introduction

Shanxi Province, as an important energy base in China, has long faced severe air pollution issues. Due to its industrial structure being dominated by high-pollution industries such as coal and steel, coupled with the influence of geographical environment and meteorological conditions, air pollution problems are particularly prominent. In recent years, although Shanxi Province has taken a series of measures for atmospheric pollution control, there are still many challenges in improving air quality. According to relevant data, the Air Quality Index (AQI) in some cities of Shanxi Province remains at moderate or even severe pollution levels during certain periods. Therefore, in-depth research on the characteristics of air quality in Shanxi Province, the influencing factors, and their changing trends is of great practical significance for formulating scientific and reasonable environmental policies and improving air quality.

This study aims to comprehensively use a variety of analysis methods to conduct in-depth analysis and prediction of air quality data in all cities of Shanxi Province from 2013 to 2024. To reveal the characteristics of air quality: through the analysis of air Quality Index (AQI) and related pollutant concentration data, reveal the spatial and temporal distribution characteristics and changing trend of air quality in Shanxi Province.

In recent years, scholars both domestically and internationally have proposed various methods in the field of air quality assessment and prediction, forming a rich theoretical system. Lin et al. systematically explored the application of Principal Component Analysis (PCA) in comprehensive evaluation, emphasizing the critical importance of data standardization and eigenvalue selection, providing a theoretical basis for dimensionality reduction of high-dimensional data[1]. Zhou et al. improved the method for determining the optimal number of clusters in K-means clustering algorithms, verifying the stability of cluster results through the elbow method and silhouette coefficient, laying a

technical foundation for air quality classification studies[2]. Jie et al. (2018) used an ARIMA model to predict dam safety monitoring data, demonstrating the limitations of time series models in nonlinear data but offering effective references for short-term predictions of stationary sequences[3]. In terms of model fusion, Si et al. systematically expounded the theoretical framework of projection pursuit models, which achieve nonlinear feature extraction through high-dimensional data projection, providing new insights for analyzing complex environmental data[4]. However, existing research often focuses on the application of single models; for example, Li et al. (1999) used PCA-based air quality evaluation, lacking dynamic predictive capabilities[5]. Meanwhile, Li Xiaoran and Xie Zhongyang's (2022) fuzzy comprehensive evaluation method, while capable of handling uncertainties, struggles to capture the evolution patterns of spatiotemporal features[6]. This study innovatively integrates the projection pursuit model with the ARIMA time series model to construct a multi-dimensional analysis framework. The projection pursuit model is used to extract the principal component features of air quality, while the ARIMA (1,0,1) model is introduced to capture the fluctuation patterns of AQI. Compared to single models, the multi-model integration framework demonstrates superior advantages in data dimensionality reduction, classification accuracy, and prediction stability.

To sum up, this study effectively reveals the spatial and temporal heterogeneity of air quality in Shanxi Province by integrating classical methods and innovative models, which provides a scientific reference for multi-source data driven environmental management decision-making, and also contributes a new perspective to the theoretical development of complex system modeling.

2. Method

Based on these theories, the projection pursuit model is established. The projection pursuit method (PP) is an exploratory data analysis technique directly driven by sample

data. It projects high-dimensional data into a low-dimensional subspace $\{z(i)\}$ through some combination for clustering, studying the structural characteristics of high-dimensional data by analyzing the variation features of data points in the low-dimensional space. Clustering involves grouping objects into sets according to certain rules or criteria, with the resulting distribution being unknown before clustering. It aims to find a one-dimensional projection direction where clusters have higher density (high cohesion) and inter-cluster distances are relatively large (low coupling), and classifies samples comprehensively based on their corresponding

composite projection values[6].

Using SPSS software, the correlation coefficient matrix is calculated. The eigenvalues and eigenvectors of the correlation matrix R are obtained, and the eigenvalues of the correlation coefficient matrix R are computed along with their corresponding eigenvectors. New indicator variables are formed from these eigenvectors, and the sample data is imported into SPSS for factor analysis. The contribution rates of the principal components of the sample are shown in the table 1.

Table 1 Factor analysis table

		Total	Variance percentage	Accumulate%
Initial eigenvalues	1	4.608	51.204	51.204
	2	1.379	15.322	66.526
	3	1.059	11.772	78.298
	4	0.741	8.229	86.527
	5	0.434	4.823	91.35
	6	0.318	3.538	94.888
	7	0.246	2.728	97.616
	8	0.13	1.45	99.066
	9	0.084	0.934	100
		Total	Variance percentage	Accumulate%
Extract the squared load	1	4.608	51.204	51.204
	2	1.379	15.322	66.526
	3	1.059	11.772	78.298
		Total	Variance percentage	Accumulate%
Square sum of repeated loads	1	3.386	37.624	37.624
	2	2.294	25.494	63.118
	3	1.366	15.18	78.298

The main component table of the sample data was obtained from the above table, and three main components were extracted. The sum of the squared loadings was 196.028%, and the sum of the squared loadings after rotation was 179.04%. The sample component matrix and the rotated component matrix are shown in the following table 2.

Table 2 Sample composition matrix

Component matrix a			
	1	2	3
PM2.5	0.922	0.187	0.12
PM10	0.874	0.195	0.217
NO2	0.832	-0.107	-0.238
SO2	0.815	-0.177	-0.368
CO	0.723	-0.081	0.469
AQI	0.642	0.614	0.033
O3_8h	-0.632	0.446	-0.312
month	-0.093	-0.71	0.356
year	-0.543	0.422	0.601

From the above chart: Principal Component 1 (PC1): In PC1, the coefficients of PM2.5, PM10, NO₂, SO₂, CO, and AQI are high and positive, indicating strong correlations among these variables in PC1. This suggests that PC1 primarily reflects comprehensive information related to air pollutant concentrations (especially particulate matter and gaseous pollutants). For example, the coefficients for PM2.5 and PM10 are 0.922 and 0.874, respectively, indicating their significant contribution to PC1; Principal Component 2 (PC2): In PC2, the coefficient for O₃ (8-hour average concentration) is positive (0.446), while the coefficients for other pollutants are negative or close to zero. This suggests that PC2 may primarily reflect independent information related to ozone

concentration, with weaker associations with other pollutants; Principal Component 3 (PC3): In PC3, the coefficient for SO₂ is negative (-0.368), while the coefficient for AQI is positive (0.033), with other variables having coefficients between these two. This may indicate that PC3 reflects differential information about certain pollutants (such as SO₂) compared to other variables.

In order to improve the interpretability of PCA and avoid result error, we adopt the caesar normalization maximum variance method. The score matrix of components after rotation is shown in the following table 3.

Table 3 Component score matrix

The rotated component matrix a			
	1	2	3
PM10	0.86	0.273	0.186
PM2.5	0.843	0.373	0.224
CO	0.838	0.115	-0.184
AQI	0.634	0.104	0.614
O3_8h	-0.618	-0.319	0.46
year	-0.032	-0.91	0.074
SO2	0.422	0.803	0.083
NO2	0.521	0.693	0.095
month	0.015	-0.027	-0.799

Explanation: Component 1 is primarily associated with atmospheric pollutants (particulate matter, gaseous pollutants). The high load values of these variables indicate their dominant role in Component 1. Component 1 may represent the overall level of air pollution, particularly the combined impact of particulate matter and gaseous pollutants; in Component 2, the load values for O₃_8h and years are negative, while those for SO₂ and NO₂ are positive. This

suggests that Component 2 is related to time trends (years) and certain specific pollutants (such as SO₂ and NO₂) with seasonal or long-term variations. The load of O₃_8h may indicate a trend opposite to that of other pollutants, which aligns with the ozone formation mechanism (usually increasing on clear days); Component 3 is mainly associated with AQI (Air Quality Index) and months. A high load of AQI indicates its significant role in Component 3, while the load of months may reflect seasonal changes in air quality. Component 3 may represent short-term fluctuations in air quality, especially related to seasonal factors (such as winter heating, summer ozone formation).

After rotation, the X values corresponding to the five principal components can be obtained, which is convenient for better derivation of factor score equation and verification of data accuracy.

Based on the analysis of the first three principal components, the eigenvalues are analyzed based on the three principal components. Based on the three principal components, this paper establishes the optimal model solution, and carries out subcategory division through the most representative eigenvalues obtained.

This paper discusses the projection pursuit of the above three principal components, further reducing the dimensionality of the sample data. Here, i represents the indicator of the i sample data. Linear dimensionality reduction is performed using the five sample data from the principal components, transforming the multidimensional data into a single-dimensional projection value with as the projection direction.

$$z_i = \sum_{j=1}^n a_j \times x_{i,j} \quad (1)$$

The structure characteristic function is constructed according to the one-dimensional distribution graph. When

the comprehensive projection value is required, the distribution characteristics of the projection value should have the characteristics of high cohesion and low coupling. Accordingly, the projection characteristic function can be constructed as

$$Q(a) = \sigma_z \times \rho_z \quad (2)$$

The standard deviation of the projection eigenvalues of the sample sequence is calculated as follows:

$$\sigma_z = \sqrt{\frac{\sum_{i=1}^n ((z_i - \bar{z}))^2}{n-1}} \quad (3)$$

$\bar{z}$ is the mean of the projection eigenvalues of all samples in the projection direction. According to the ideas of probability theory and mathematical statistics, The larger the σ_z , The lower the coupling degree of the sample set, the greater the dispersion between classes. Let the difference between the projection values of any two samples k and l be:

$$r_{kl} = |z_k - z_l| \quad (4)$$

The local density p_z of the projection value can be calculated by the following formula:

$$p_z = \sum_{k=1}^m \sum_{l=1}^m (R - r_{kl}) \mu(R - r_{kl}) \quad (5)$$

where R is the density window width parameter, and the optimal value of the density window width parameter is:

$$R = \frac{\max(r_{kl}) + n}{2} \quad (6)$$

For the projection indicator function. When the sample sets of each indicator value are given, the projection indicator function only changes with the variation in the projection direction α . The optimal projection direction that maximizes the projection indicator function can be estimated to best reveal certain feature structures of high-dimensional data. After cleaning and processing the data, a visualization was performed, resulting in an air quality index map for Shanxi from 2013 to 2024, as shown in Figure 1:

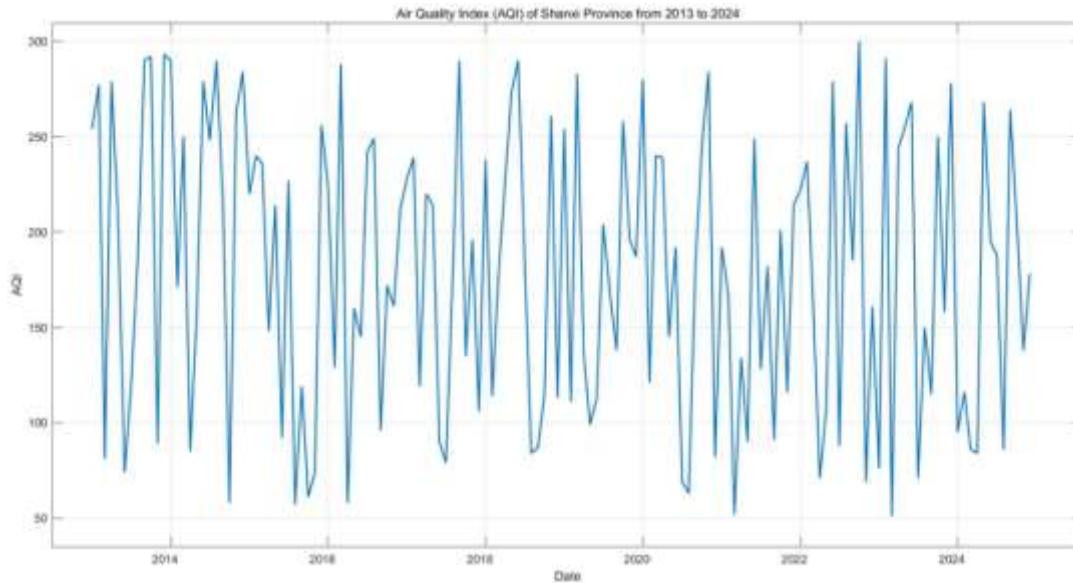


Figure 1 Shanxi Air Quality Index from 2013 to 2020

As shown in the figure above, the data exhibit significant fluctuations and may contain seasonal trends. Moreover, the AQI values fluctuated considerably between 2013 and 2024, with no clear linear trend. For stationarity testing, an ADF test was used. The P-value of the original data is 0.0038, which is

less than the significance level of 0.05, indicating that the original data is stationary. Therefore, the difference order $d=0$, and the ARIMA parameters (p, d, q) are determined as shown in Figure 2:

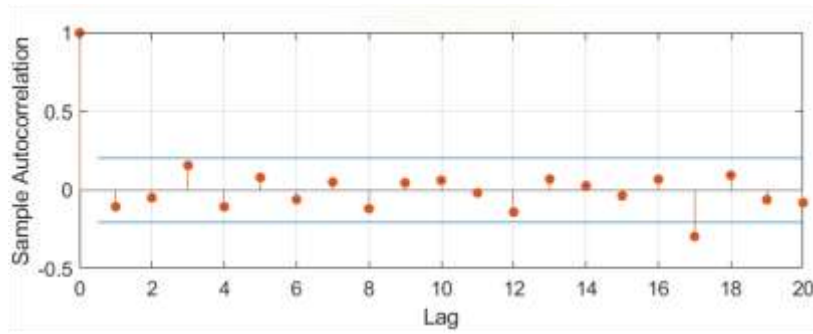


Figure 2 ACF diagram

It can be obtained from ACF that the truncation at lag 1 indicates $q=1$, as shown in Figure 3.

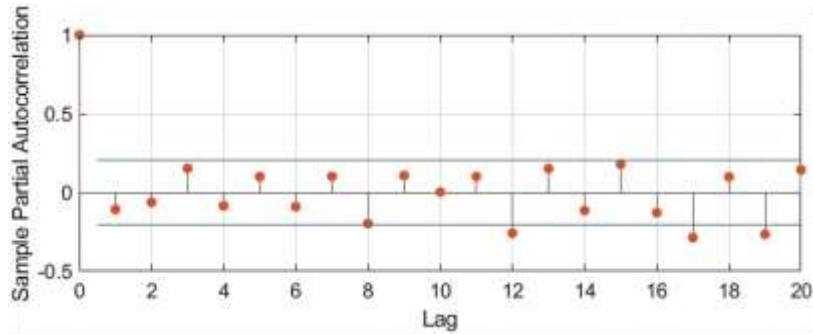


Figure 3 PACF diagram

According to the PACF diagram, truncation at lag 1 indicates that $p=1$. Therefore, the ARIMR (1,0,1) model is selected for synthesis.

3. Results

The predicted results are shown in Figure 4.

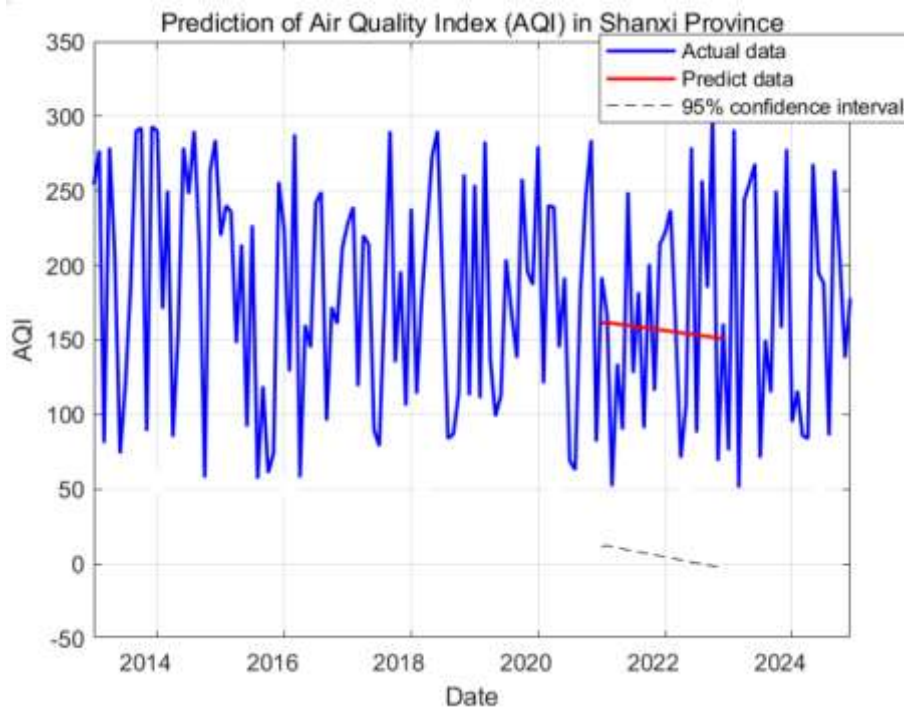


Figure 4 Prediction of Air Quality Index in Shanxi Province

The results fluctuate between 100 and 200, and the 95% confidence interval provides a range of uncertainty for the prediction. Overall, the ARIMA(1,0,1) model has shown some effectiveness in predicting air quality, but there is still room for improvement. The projection pursuit model has processed the influencing factors of air quality from high

latitude to low latitude, which to some extent makes the air quality closer to the actual air quality.

The ARIRM (1,0,1) model can fit the AQI data of Shanxi Province from 2013 to 2024 well. The residual is white noise, indicating that the model captures the main features of the data. The predicted results provide a reasonable estimate of

future AQI.

4. Conclusions

This study focuses on the air quality data of various cities in Shanxi Province from 2013 to 2024, employing a comprehensive set of analytical methods to comprehensively evaluate and predict air quality. The research first uses SPSS software to conduct KMO and Bartlett tests to analyze data applicability. After standardization, factor analysis is applied to extract key principal components and construct a factor score equation. Innovatively, the Matlab time series toolbox is utilized for rigorous stationarity tests and parameter optimization, establishing an ARIMA (1,0,1) prediction model, which performs well in both fitting and predictive accuracy. Additionally, the projection pursuit model is used to reduce the dimensionality of factors influencing air quality, further uncovering data characteristics. The results show that the multi-model fusion analysis framework can effectively reveal the spatiotemporal features and influencing factors of air quality in Shanxi Province. This study not only provides a scientific basis for dynamic monitoring and management of air quality in Shanxi Province but also offers new perspectives and ideas for theoretical development in environmental management decisions driven by multi-source data and

complex system modeling.

References

- [1] Lin Haiming, Du Zifang. Issues to be paid attention to in comprehensive evaluation of principal component analysis [J]. Statistical Research, 2013,30(8):25-31.
- [2] Zhou Shibing, Xu Zhenyuan, Tang Xuqing. A new K-means algorithm for determining the optimal number of clusters [M]. Computer Engineering and Application, 2010.
- [3] Jie Jiancang, Wang Yue, Lei Sheping et al. Data Analysis and Prediction of Dam Safety Monitoring Based on ARIMA Model [J]. People's Yellow River, 2018.
- [4] Si Shoukui, Sun Xijing. Mathematical Modeling Algorithms and Applications [M]. National Defense Industry Press, 2011.
- [5] Li Yanshuang, Zeng Zhenxiang, Zhang Min et al. Application of principal component analysis in multi-indicator comprehensive evaluation [J] Journal of Beijing University of Technology, 1999.
- [6] Li Xiaoran, Xie Zhongyang. Research on Construction Project Risk Evaluation Based on Entropy Weight-Fuzzy Comprehensive Evaluation Method [J]. Safety and Health, 2022(06):58-62.