World Scientific Research Journal Volume 11 Issue 10, 2025
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

Research on Functional Data Method Based on Air Quality Index of Sichuan Basin

Chengbing Zhao, Ruimin Zhang, Yifei Wu

School of Mathematics & Physics, Anhui Jianzhu University, Hefei, China

Abstract

So far, there have been relatively few studies on air quality in the Sichuan Basin. This paper uses the functional data analysis method to study the air quality data of the Sichuan Basin. Firstly, the discrete air quality data of 21 prefecture-level cities in the Sichuan Basin were smoothed using the Fourier basis function to transform them into continuous functional data. Then, the principal component analysis of the smoothed air quality index was conducted. Secondly, the non-negative matrix factorization method is introduced to achieve data dimensionality reduction and retain key information through the linear combination of basis vectors, and to conduct cluster analysis on the air quality data of the Sichuan Basin. The research results show that the fluctuation of the air quality index in the Sichuan Basin has a distinct seasonal pattern. The cumulative variance contribution rate of the first three principal components reached 95.3%, which can explain most of the original data. According to the clustering results, the Sichuan Basin can be divided into three types of regions, namely the core area, the marginal area and the mountainous area in southwest China. These studies provide data support and theoretical basis for the future pollution prevention and control policies in the Sichuan Basin.

Keywords

Air quality; Functional principal component analysis; Non-negative matrix factorization; Cluster analysis.

1. INTRODUCTION

China's economy is developing rapidly, the urbanization process is constantly advancing, and air pollution is becoming increasingly serious. As a special geographical area, the air quality problem in the Sichuan Basin urgently needs to be solved. The Sichuan Basin is surrounded by high mountains such as the Daba Mountains, Longmenshan Mountains and Minshan Mountains, forming a closed climate and environment. Due to the frequent occurrence of temperature inversion, pollutants are difficult to diffuse, resulting in significant fluctuations in air quality and relatively long retention times of pollutants. The energy structure in the Sichuan Basin is dominated by coal, heavy industry is constantly developing, and there are also traffic emissions, etc. All these are the main factors causing air pollution. In response to this situation, the Sichuan Provincial Government has issued relevant policies on pollution prevention and control to promote pollution control efforts, optimize the energy structure, and strengthen the construction of environmental monitoring facilities. However, due to the influence of various factors such as geography, climate and industrial structure, governance work still faces many challenges.

Functional data analysis is a method of processing discrete data into continuous functional data[1]. This method has been widely applied in multiple fields such as meteorology, environment and medicine. For instance, it can be used to convert discrete observations of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

precipitation into functional data to simulate precipitation patterns in different regions of Malaysia[2]. In addition, this method is also applied to medical data modeling[3], vehicle-mounted GPS signal processing[4], and the spatio-temporal evolution analysis of air pollution[5]. In China, existing studies have utilized functional principal component analysis methods to investigate the characteristics of air pollution in the Fen-Wei Plain[6], regional water usage changes[7], and to review and summarize the FDA regression methods[8]. These studies show that the FDA can transform discrete observational data of air quality into smooth function curves, more accurately depicting the dynamic changes of pollutants in time and space.

Functional cluster analysis is a clustering method that further processes functional data on the basis of functional data analysis. It achieves clustering and grouping by defining similarity measures between functions or using statistical models, and is widely applied in fields such as medicine and health sciences, and environmental sciences. Relevant research includes cluster analysis of the employed population data in prefecture-level cities of Guangdong Province[9], time series characteristic analysis of air quality in the Yangtze River Delta[10], and the proposal of a clustering method combining principal component analysis and functional distance measurement optimization[11], which effectively enhances the accuracy and practicality of clustering.

2. ANALYSIS OF AIR QUALITY CHARACTERISTICS IN THE SICHUAN BASIN

2.1. Introduction to Data Sources

This paper selects the air quality index of some cities in Chongqing, Sichuan Province, Guizhou Province and Yunnan Province, which are covered by the Sichuan Basin, as the research object, and conducts a systematic analysis of the changing law of air environmental quality. AQI refers to the dimensionless relative value of the degree of air pollution or air quality grade. It is a new "Ambient Air Quality Standard" issued by the state in March 2012, used to describe the air quality situation, as shown in Table 1. The larger the air quality index value, the higher the level and category, indicating a more severe degree of air pollution and a greater threat to people's health.

Table 1. Air Quality Grading Standards

AQI value	AQI level	AQI category
0-50	Level 1	Optimal
51-100	Level 2	Good
101-150	Level 3	Mild pollution
151-200	Level 4	Moderate pollution
201-300	Level 5	Serious pollution
301-500	Level 6	Heavy pollution

This paper selects the average monthly AQI concentration values from January 1, 2019 to November 30, 2023 as the research object, and all the data are sourced from the China National Environmental Monitoring Centre. The data is genuine, reliable and timely, which enables the research to proceed smoothly.

2.2. Experimental Section

Data in numerous fields such as economics, finance, meteorology, and medicine are often rather complex, existing in the form of a combination of time and space. When confronted with more complex problems, traditional multi-method approaches can no longer meet the demands, and new research methods need to be developed. Functional data emerged as The Times

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

require, and its development has become inevitable. Since the Canadian statistician Ramsay mentioned the new concept of functional data in 1982, functional data has attracted great attention in the industry and gradually formed diverse analytical methods.

The storage form of functional data is shown in Table 2 Suppose there are n samples of functional data.

Table 2. Functional data storage format

	X1	X2	...	Xn
t1	y11	y21	...	yn1
t2	y12	y22	...	yn2
...	...	...	...	...
tN	y1N	y2N	...	ynN

Suppose the collected discrete data $y = (y_1, y_2, \cdots, y_n)$ is used to establish a mathematical model as follows:

$$y_i = x_i(t) + \varepsilon_i, i = 1,2, \cdots, n, \tag{1}$$

Among them, $x_i(t)$ is the function curve to be fitted,and ε_i is the error in the original data.

The concept of basis functions is: in a system of functions Φ_k composed of a series of functions φ_k. The functions in the series are independent of each other, and any function can be represented in the form of a linear combination of basis functions. The basis function smoothing method is to perform fitting by means of the linear combination of K known basis functions in the basis function system. Specifically, If A linear combination of a set of basis functions $\Phi(t) = \{\varphi_1(t), \varphi_2(t), \cdots, \varphi_k(t)\}$ is selected, then there will exist coefficient vectors $\xi^T = (\xi_1, \xi_2, \cdots \xi_k)$ such that:

$$x_i(t) = \sum_K^{k=1} \xi_k \varphi_k(t) = \xi^T \Phi(t), \tag{2}$$

Among them, $\varphi_k(t)$is the Kth basis function, and $\xi_k = \int x_i(t)\varphi_k(t)dt$ is the projection of $\varphi_k(t)$ on $x_i(t)$.

Collect the monthly average air quality index data of 21 prefecture-level cities in the Sichuan Basin from 2019 to 2023. The functional data method was selected to preprocess the data. The change in time was taken as the horizontal axis, and the average monthly AQI concentration value was taken as the vertical axis to draw the initial data distribution graph, as shown in Figure 1.

As can be seen from Figure 1, the AQI data of 21 cities in the Sichuan Basin from 2019 to 2023 have a certain periodicity. Therefore, in this paper, the Fourier basis function is adopted to fit the original data, transforming the discrete data of the average monthly AQI concentration values of 21 cities into continuous function curves. To enhance the smoothness of the function curve, we introduce A penalty term $\lambda \bullet PEN_L(x)$. At this point, assuming $x(t)$ is differentiable, the mean square error with the penalty term can be defined as:

$$PENSSE_\lambda(x|y) = \sum_{j=1}^N [y_j - x(t_j)]^2 + \lambda \times PEN_L(x) \tag{3}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

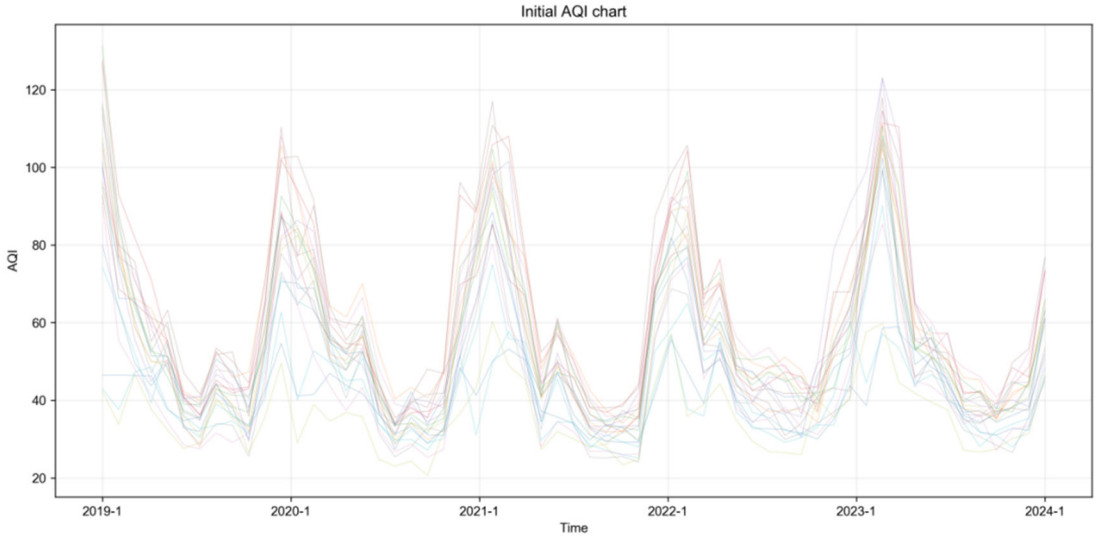

Figure 1. Original distribution map of air quality index in Sichuan Basin from 2019 to 2023

Here, $PEN_L(x) = \int [(Lx)^2](t)dt = ||Lx||^2$ is the penalty term, λ is the penalty parameter, and L is the linear differential operator, which can be taken as: $Lx(t) = \omega^2 D + D^3$. The fitting curve after adding the penalty term is shown in Figure 2.

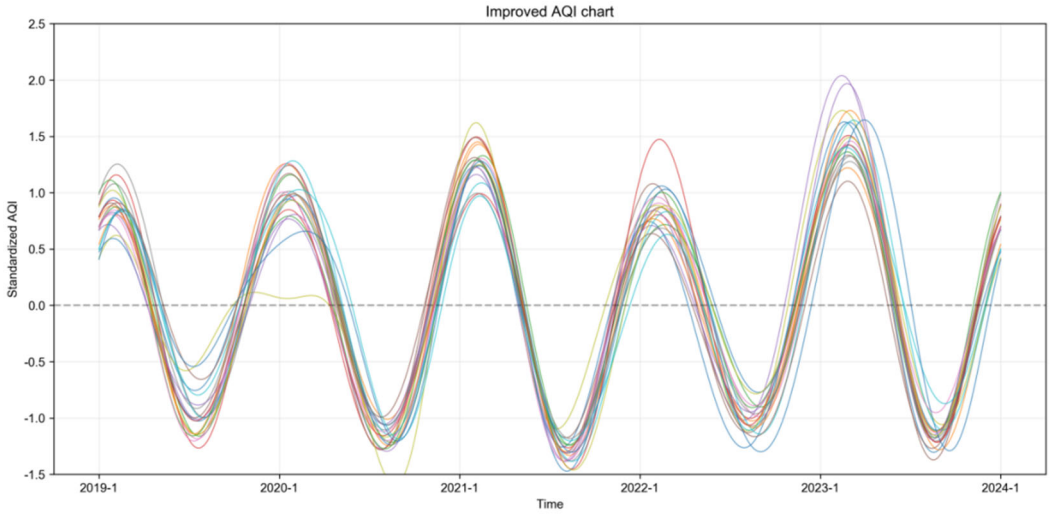

Figure 2. Fitting Chart of Air Quality Index in Sichuan Basin from 2019 to 2023

As can be seen from Figure 2, the Fourier basis function fitting function curve not only retains the vast majority of the characteristics of the original data, but also shows a certain degree of smoothness, thus transforming the originally discrete observed data into functional data with continuous features, which is convenient for further analysis and research.

As can be seen from Figure 2, during the period from 2019 to 2023, the Air Quality Index (AQI) of 21 prefecture-level cities in the Sichuan Basin showed significant fluctuations. The AQI curve peaks in winter and troughs in summer, showing a significant seasonal fluctuation pattern. That is to say, air quality deteriorates severely in winter and improves in summer. This situation is mainly influenced by winter heating, increased industrial emissions, and special phenomena in the Sichuan Basin (stable weather and temperature inversion). In summer, the enhanced precipitation and air mobility are conducive to the diffusion and removal of pollutants.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

Although the changing trends in various cities are generally the same, there are still differences in the peak values in winter and the trough values in summer. The peak pollution in some cities during winter is significantly higher than that in others, indicating that these cities are facing more severe pollution problems in winter. Meanwhile, some cities may experience relatively high AQI values in certain years, which might be related to the local severe pollution weather conditions (smog). In addition, the overall trend indicates that during the period from 2019 to 2023, the AQI peaks in some cities may have declined. This suggests that as time goes by, the measures for air quality control may have begun to show certain results, but the overall fluctuation trend remains stable. The characteristics of air quality in the Sichuan Basin are quite obvious, and the pollution problem in winter is particularly prominent. The main differences among various cities are reflected in the AQI values in winter, while in summer, the air quality is relatively close.

3. FUNCTIONAL PRINCIPAL COMPONENT EMPIRICAL ANALYSIS OF AIR POLLUTION CHARACTERISTICS IN THE SICHUAN BASIN FROM 2019 TO 2023

Based on the periodic characteristics of the monthly average AQI concentration data collected from 21 cities in the Sichuan Basin over the time scale from 2019 to 2023, this section selects the Fourier basis function as the functional processing of the original discrete data. Then, it reduces the dimension of the monthly average AQI concentration values of the 21 cities in the Sichuan Basin from 2019 to 2023 and extracts multiple principal components. Analyze the influence of the influencing factors represented by the principal components on the monthly average AQI concentration value.

As shown in Table 3, when the principal components of the first three functions are taken, the cumulative variance contribution rate reaches 95.30%, which is far greater than 85%, and can explain the vast majority of variables.

Table 3. Table of eigenvalues and cumulative contribution rates of three principal components

	The first principal component	The second principal component	The third principal component
Characteristic value	19.0943	0.8983	0.3646
Variance contribution rate	89.38%	4.21%	1.71%
Cumulative variance contribution rate	89.38%	93.59%	95.30%

Table 3 also presents the results of principal component analysis, where the eigenvalue of the first principal component is 19.0943, and the variance contribution rate is as high as 89.38%, indicating that most of the original data can be explained by the first principal component, which is significantly representative. The eigenvalues of the second and third principal components were 0.8983 and 0.3646 respectively, and their explanatory power for data variation was relatively weak, with variance contribution rates of only 4.21% and 1.71% respectively. The cumulative variance contribution rate of the first three principal components reached 95.30%, indicating that retaining these three principal components can well preserve the main information in the original data, achieve effective dimensionality reduction of high-dimensional data, and not only reduce the data dimension but also maintain the integrity of the information.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

Overall, the data in Table 3 shows that there are significant differences in the main factors of air quality among different cities. Climatic conditions are associated with the first principal component. The differences in heating and emissions are related to the second principal component. Humidity is related to the third principal component. The cumulative contribution rate of the first three principal components was 95.30%, which could explain the changes in the vast majority of the original data. This indicates that the analysis based on these three principal components can effectively summarize the characteristics and main factors of the monthly average AQI concentration value changes in 21 cities in the Sichuan Basin, thereby providing reliable data basis and theoretical support for subsequent research.

4. NON-NEGATIVE MATRIX FACTORIZATION CLUSTERING ANALYSIS RESULTS OF AIR QUALITY INDEX IN SICHUAN BASIN FROM 2019 TO 2023

The average monthly AQI index of 21 prefecture-level cities in the Sichuan Basin from 2019 to 2023 was analyzed through the non-negative matrix factorization clustering method. The results are shown in Figure 3 and Figure 4.

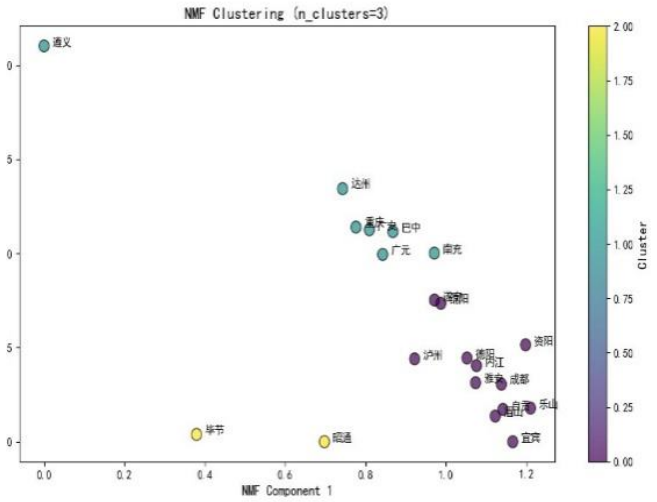

Figure 3. Non negative Matrix Decomposition Cluster Diagram of Annual Average AQI Concentration in Sichuan Basin from 2019 to 2023

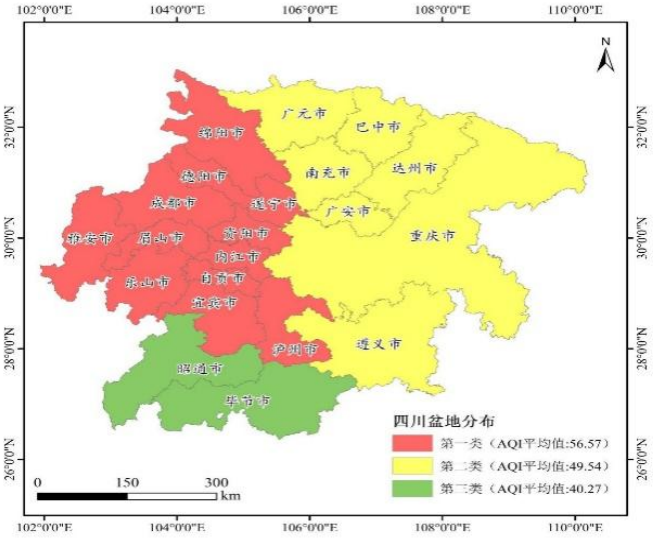

Figure 4. Spatial distribution map of Sichuan Basin from 2019 to 2023

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

The results of clustering are classified into three categories. The first category includes 12 cities: Chengdu, Mianyang, Yibin, Luzhou, Zigong, Deyang, Suining, Neijiang, Leshan, Meishan, Ya 'an and Ziyang. The air quality in these cities is the worst. These areas belong to the economic, industrial and transportation centers of Sichuan Province. The relatively high AQI value indicates that the air pollution here is rather severe. As the capital of Sichuan Province, Chengdu has a highly concentrated population and enterprises and serves as the economic hub of the entire province. Heavy traffic flow and frequent industrial activities, these high-intensity human activities are the main sources of pollution. Cities such as Mianyang and Deyang are dominated by manufacturing and heavy industry. These industrial activities generate large amounts of sulfur dioxide (SO2), nitrogen oxides (NO youdaoplaceholder6) and particulate matter (PM2.5 and PM10). The second category includes seven cities: Chongqing, Nanchong, Zunyi, Guangyuan, Guang 'an, Dazhou and Bazhong, whose air quality is at a medium level. These regions are also areas with relatively high population density and economic activity intensity. However, compared with the first category of areas, their industrial density and pollution emission intensity are relatively lower, and thus their environmental quality is relatively better. Although these regions do not have as high an industrial and population density as the first category of areas, they remain important economic regions. Transportation and domestic pollution sources still have a certain impact on environmental quality. Chongqing is rather special. Although it is adjacent to the southeastern edge of the Sichuan Basin, as a municipality directly under the Central Government, its urbanization level and economic development degree are similar to those of the first category of regions. The third category includes the two cities of Bijie and Zhaotong. The air quality here is the best and most stable. Bijie and Zhaotong are less affected by pollution and are the areas with the cleanest air. The degree of industrialization and urbanization in these regions is relatively low, the pressure on the environment caused by human activities is small, and the sources of pollution emissions are limited. These areas are located at or close to the edge of the basin, with good air circulation conditions, which are conducive to the diffusion of pollutants. Therefore, the natural environment has been well protected.

5. RESULT ANALYSIS

5.1. Conclusions

This paper uses Fourier basis functions to fit the average monthly AQI data of 21 prefecture-level cities in the Sichuan Basin from 2019 to 2023. This method transforms the originally discrete data into smooth functional curves, not only retaining the original characteristics of the data but also demonstrating strong continuity. The results show that AQI has a distinct fluctuation pattern: in winter, due to the influence of factors such as heating, industrial emissions, and temperature inversion, AQI reaches its peak. In summer, due to the increase in precipitation and the enhancement of air mobility, the AQI improves significantly. Although the trends in various cities are generally the same, there are differences between the peak values in winter and the trough values in summer. In some cities, pollution is more severe in winter, and the high AQI that occurs in individual years may be related to local heavy pollution weather. Overall, the peak of winter in some cities from 2019 to 2023 has slightly declined, indicating that the governance measures have achieved certain results. However, the overall fluctuation trend remains relatively stable.

Based on the monthly average AQI concentration data of 21 cities in the Sichuan Basin from 2019 to 2023, the Fourier basis function was used for functional processing and dimension reduction operations, and then the top three principal components were extracted. The cumulative contribution rate of these three principal components can reach 95.3%, retaining most of the original data. The degree to which the changes in air quality in different

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

administrative regions are influenced by principal components varies. In some cities, the improvement of air quality is dominated by a single principal component, while in others, it is affected by the combined effect of multiple factors. This indicates that the main driving factors for air quality changes in the Sichuan Basin include climatic conditions, heating emissions, and humidity, etc. Principal component analysis provides important support for analyzing the characteristics of urban air quality.

Based on the analysis of the data of the Sichuan Basin from 2019 to 2023, this area is classified into three categories. The environmental quality of the first category is the worst, mainly concentrated in the core areas of the basin such as Chengdu, Mianyang, Deyang, Meishan and Leshan. These prefecture-level cities are the centers of economy, industry and transportation. Due to the influence of high-intensity human activities and industrial pollution, the air pollution situation is very serious. The environmental quality of the second category is relatively poor, and it is distributed at the edge of basins or in places with less dense transportation such as Chongqing, Dazhou, Guangyuan and Guang 'an. The industrial density in these places is relatively low, but they are still affected by economic activities and domestic pollution. The environmental quality of the third category is relatively the best. In mountainous areas such as Bijie and Zhaotong, the degree of industrialization is relatively low, the air circulation conditions are better, and the natural environment has been better protected.

In this paper, the Fourier basis function smoothing method is employed to fit discrete data into continuous functional data, which provides convenience for subsequent analysis. Based on the fitting function, the data was dimensionally reduced through principal component analysis. The first three principal components were selected, and their cumulative variance contribution rate reached 95.30%, which could represent most of the data. After conducting cluster analysis on 21 prefecture-level cities in the Sichuan Basin, the obtained results are highly consistent with the actual situation.

5.2. Suggestions

To improve the air quality in the Sichuan Basin, efforts should be made in multiple aspects. In terms of industrial layout, first of all, it is necessary to promote the transformation of highly polluting enterprises, popularize clean energy, reduce the consumption of fossil fuels, increase the proportion of renewable energy, implement green manufacturing and clean production technologies, and reduce wastewater discharge, etc. Secondly, it is necessary to reduce the proportion of highly polluting industries in the core economic zone, transfer some industries to the surrounding areas, replace traditional energy with clean energy to develop green industries, and optimize the industrial structure. In this way, air quality can be improved from the source. In addition, efforts are also needed in the field of transportation. On the one hand, it is necessary to promote new energy vehicles and enhance the appeal of public transportation, thereby reducing the proportion of private car trips and lowering exhaust emissions. On the other hand, strict standards should be set for high-emission vehicles and traffic restriction policies should be implemented. For instance, cities like Luzhou and Yibin have been promoting the use of clean energy vessels to control water transport exhaust. At the same time, increasing investment in the monitoring and control of pollution sources is also extremely crucial. By leveraging technological means to conduct real-time monitoring and control of pollution sources, promoting the upgrading and transformation of industrial enterprises, and adopting environmentally friendly process equipment to reduce emissions, both environmental and economic benefits can be achieved. Due to the impact of winter temperature inversion on air quality, it is necessary to enhance monitoring and response, issue early warnings and take emergency measures, such as restricting industrial emissions, to ensure public health and environmental stability.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0005

Through the implementation of the above comprehensive measures, the air quality in the Sichuan Basin will gradually improve, and the AQI value will also decrease accordingly, promoting the ecological environment of the entire region to develop in a better direction.

ACKNOWLEDGMENTS

This paper was supported by Key Project of Natural Science Foundation of Higher Education Institutions in Anhui Province (KJ2021A06312024AH050257) and Humanities and Social Sciences Fund Project for Higher Education Institutions in Anhui Province (2023AH040035).

REFERENCES

[1] J.O. Ramsay: When the data are functions, Psychometrika, Vol. 47 (1982) No. 4, p.379-396.

[2] J. Suhaila, A.A. Jemain, M.F. Hamdan and W.Z.W. Zin: Comparing rainfall patterns between regions in Peninsular Malaysia via a functional data analysis technique, Journal of Hydrology, Vol. 411 (2011) No. 3-4, p.197-206.

[3] H. Sørensen, J. Goldsmith and L.M. Sangalli: An introduction with medical applications to functional data analysis, Statistics in Medicine, Vol. 32 (2013) No. 30, p.5222-5240.

[4] Y. Méneroux, A. Le Guilcher, G. Saint Pierre, M. Ghasemi Hamed, S. Mustière and O. Orfila: Traffic signal detection from in-vehicle GPS speed profiles using functional data analysis and machine learning, International Journal of Data Science and Analytics, Vol. 10 (2020) No. 1, p.101-119.

[5] B.M. Petronio, M. Pietrantonio, M. Pietroletti and N. Cardellicchio: *Metal Speciation and Bio-Availability in Marine Sediments of Northern Adriatic Sea, Proc. Seventh FECS Conference on Environmental Science and Pollution Research* (Rome, Italy, 2000), Vol. 1, p.320.

[6] J.H. Bai: *Functional Data Analysis of Air Quality in Fen-Wei Plain* (MS., Guangxi Normal University, China 2023), p.15.

[7] M. Li: *Analysis of water consumption data function based on principal component analysis method* (MS., Hefei University of Technology, China 2015), p.10.

[8] D.I.N.G. Hui, W.C. Xu, H.B.U. Zhu, G.W.A.N.G. Guochang, Z.H.A.N.G. Tao and Z.H.A.N.G. Riquan: Review of Regression Analysis for Functional Data, Chinese Journal of Applied Probability and Statistics, Vol. 34 (2018) No. 6, p.630-654. (In Chinese)

[9] W.P. Wang: *Based on the Theory of Functional Data to Classify the Working Population* (MS., Northeast Normal University, China 2010), p.10.

[10] X.J. Cheng: *Functional Data Analysis of Air Quality in Yangtze River Delta* (MS., Guangxi Normal University, China 2021), p.20.

[11] Y. Zhang: *Research on Functional Data Analysis Method Based on Runoff Data of the Main Stream of the Yellow River* (MS., Lanzhou University of Finance and Economics, China 2024), p.22.