Air Quality Analysis of Wuhan from the Perspective of Functional Data

Chen Guici¹, Zhang Zuo* and Zuo Qian²

¹Hubei Province Key Laboratory of System Science in Metallurgical Process, Wuhan University of Science and Technology, Wuhan, Hubei, 430065, P.R. China
²Hubei University of Technology, Wuhan, Hubei, 430068, P.R. China
E-mail: chenguici@wust.edu.cn, *571137033@qq.com, 779110720@qq.com

Abstract. In recent years, with the continuous advancement of Wuhan's industry, manufacturing and engineering construction, the consumption of energy is also increasing, and the air quality problems caused by it have also received attention from all walks of life. In order to explore the characteristics of air pollution in Wuhan, this paper uses functional data analysis methods to study air quality (AQI) from the perspective of space and time. The results show that 13 prefecture-level cities in Wuhan can be divided into 4 levels according to the quality of air quality. In general, Xinzhou District, Huangpi District, Jiangxia District and Hannan District have the best air quality, while Qingshan District and Dongxihu District have the worst air quality; all have significant seasonal characteristics in the time dimension, that is, winter is a high-frequency band with serious air pollution.

1. Introduction
In recent years, with the advancement of China's industry, agriculture and chemical and manufacturing industries, the rapid growth of the population and the rapid improvement of the urban economy have become an obvious trend, but the consumption of energy is also increasing, which has caused environmental quality problems. It has increasingly become the focus of attention from all walks of life, and air quality is one of the environmental protection focuses that people are most concerned about. China also formally implemented the "Environmental Protection Tax Law of the People's Republic of China" in January 2018, and at the Third Plenary Session of the first session of the 13th National People's Congress held in Beijing in the same year, "ecological civilization" was officially incorporated into the constitution.

Air quality affects people's health to a great extent and is one of the important guarantees for sustainable economic development. Wuhan is the capital of Hubei Province, an important industrial, scientific and educational base and a highly comprehensive transportation hub in the country. It has the alias of nine provinces and is also an area with significant changes in air quality in China. The number of days with excellent air quality decreased from 69.86% in 2017 to 68.22%, but it increased by 3.29% compared with 2016. Among the 106 pollution days in 2018, the main pollutants were mainly fine particles (PM₂.₅) and ozone (O₃).

For the study of air quality, a large number of scholars start from the perspective of environmental science and meteorology [1-3], such as Cheam, A. S. M [4] et al. proposed a new hybrid clustering method for spatiotemporal data (STM), and proved the identifiability of the model. Considering that air quality prediction is a challenging problem because it relies on many factors such as weather factors and spatial
and temporal differences, Junshan Wang et al\[5\], designed a data-driven method. Experimental results show that each component of the model effectively improves the prediction accuracy. Based on the Beijing AQI data from 2013 to 2017, Deng found that the spring air quality index is the highest and the summer air quality index is the lowest, showing a clear seasonal effect. Based on the Beijing AQI data from 2013 to 2017, Tian Y found that the spring air quality index is the highest and the summer air quality index is the lowest, showing a clear seasonal effect\[6\]. Taking the city of Wuhan as an example, Gang Xu et al. discussed the quantitative relationship between land use and air quality based on the data from nine monitoring points in 2007-2014. The results show that land use has a significant impact on air quality\[7\].

It can be seen from the above literature analysis that when analyzing and studying the main influencing factors of air quality, the source of pollutants, and the spatial and temporal characteristics of air quality, the research object is mainly the data collected by different attributes at the same time or the same "Cross-section data" in time, but the data obtained in practice are often "panel data"; or the latitude of the research data is too large. Therefore, in order to break through the limitations of unequal intervals or different types of observation data, non-parametric fitting methods are used to fit discrete air quality data points into curves, and function data methods are best used to analyze the curves with the regional data of Wuhan with finer granularity as the background, to study the characteristics of air quality distribution, to capture the differences and similarities in air quality between regions, and to dig out more data information.

The remaining parts of the article are planned as follows. Section 2 describes the applied method and sample data. Section 3 presents and discusses the empirical results, and conclusion analysis is provided in Section 4.

2. Methodology and model specification

Functional data was proposed by the statistician Ramsay in the 1970s and was successfully applied to the study of economic data with time characteristics, temperature changes and height changes. Its characteristics are that it can contain infinite-dimensional data, but this Modularity and inference bring certain difficulties. Therefore, the key to modeling functional data is to reduce the dimensionality of the data features. This paper mainly uses the functional principal component analysis\[11-12\].

Clustering analysis is mainly aimed at sample observation data without prior information, grouping according to the similarity of the sample in nature, it not only can obtain the distribution of the overall data, but also for some specific categories At the same time, the clustering methods for functional data slowly matured, and they are mainly divided into three categories: non-parametric methods using specific distances or curve differences, model-based clustering, and clustering after functional dimensionality reduction analysis.

2.1. Functional principal component analysis

Suppose there is an integrable function \( x_i(s)(s = T), i = 1, 2, \ldots, n \) in the interval T, and turn it into a comprehensive variable:

\[
f_i = \int_T \beta(s)x_i(s)ds, i = 1, 2, \ldots, n.
\]

After standardization and other operations are available:

\[
\text{Var}(f) = \frac{1}{n-1} \sum_{i=1}^{n} \left( \int_T \beta(s)x_i(s)ds \right)^2.
\]

The covariance is:

\[
\nu(s,t) = \frac{1}{n-1} \sum_{i=1}^{n} x_i(t)x_i(s).
\]

The first principal component meets the following conditions:
\[ \max \frac{1}{n-1} \sum_{i=1}^{n} \left( \int_{T} \beta_i x_i \, ds \right)^2. \]

Similarly, the principal component meets the following conditions:

\[ \max \frac{1}{n-1} \sum_{i=1}^{n} \left( \int_{T} \beta_i x_i \, ds \right)^2 \]

\[ s.t. \int_{T} (\beta_j(s))^2 \, ds = \| \beta_j \|^2 = 1 \]

The score of the i-th sample curve \( x(s) \) the jth principal component is:

\[ f_{ij} = \int_{T} \beta_j(s) x_i(s) \, ds, \quad i = 1, 2, \cdots, n. \]

### 2.2. Principal components of air quality curve

Each sample curve \( x(t) \) can be approximated by a set of orthogonal basis functions \( \{ f_1, f_2, \cdots, f_L \} \) as:

\[ x_i(t) \approx \hat{x}_i(t) = \sum_{m=1}^{L} \alpha_{im} f_m(t), \quad \alpha_{im} = \int x_i(s) f_m(s) \, ds. \]

Minimize the objective function:

\[ PCASSE = \sum_{i=1}^{N} \left\| x_i - \hat{x}_i \right\|^2. \]

\[ \sum_{i=1}^{N} \left\| x_i - \hat{x}_i \right\|^2 = \int_{0}^{T} (x(s) - \hat{x}(s))^2 \, ds, \quad \alpha_{im} = \int x_i(s) f_m(s) \, ds \]

Where \( \hat{x}_i \) is the main component score.

### 3. The empirical results

#### 3.1. Data Sources

This article uses the air quality daily test data published by the Wuhan Municipal Eco-Environmental Bureau and the Wuhan Meteorological Bureau official website from January 1, 2017, to June 7, 2019. A total of 14 air quality test stations data, including urban test station data, 6 state-controlled automatic detection station data and 7 city-level automatic detection station data. A total of 881 samples were used to explore the spatial and temporal distribution characteristics of air quality in Wuhan based on the above data.

#### 3.2. Model building

Before the establishment of the model, it is necessary to convert the discrete data into functional data and use a linear combination of basis functions. In this paper, the Fourier basis is used. The expression is as follows:

\[ \Phi_0(t) = 1, \Phi_{2r-1} = \sin rwt, \Phi_{2r}(t) = \cos rwt \]

#### 3.2.1. Penalty factor determination

When fitting and smoothing function type data, the most commonly used is the non-parametric fitting method, that is, the rough punishment method, that is, the function is trimmed and smoothed by punishing it. The formula for calculating the penalty factor is as follows:
It can be polished from the above formula. We hope that the sum of squared errors is larger and that we have enough degrees of freedom to characterize the data. Therefore, replace the smaller GCV as the penalty factor of the curve, so as to minimize the loss and smooth the curve. As follows:

Therefore, in order to break through the limitations of unequal intervals or different types of observation data, non-parametric fitting methods are used to fit discrete air quality data points into curves, and function data methods are best used to analyze the curves.

\[
GCV(\lambda) = \left( \frac{n}{n - df(\lambda)} \right) \left( \frac{SSE}{n - df(\lambda)} \right)
\]

(10)

As can be seen from Figure 1, the GCV curve is a U-shaped curve. When \( \log_{10} \lambda = 10 \), the value of GCV is the smallest, i.e., \( 10^{10} \) as the penalty factor.

3.2.2. Functional principal component analysis

In this paper, the collected samples are analyzed using the functional principal component analysis (FPCA) model. The analysis results are shown in the following figure. The solid line represents the average change curve of air quality, and the "+" curve and "-" curve represent the floating range of the mean.

Figure 2 shows that the variance contribution rates of the first two principal components to air quality are 78.1% and 14.1%, respectively, and the cumulative contribution rate reaches 92.2%, which can represent most of the sample information to a certain extent. Therefore, this article replaces the first two principal components to analyze the overall change of air quality in Wuhan.

The first main component mainly explains the two periods of time at the beginning of the year and the end of the year, during which the air quality changes are more obvious, that is, the obvious winter effect is highlighted. In 2017, the air quality index category appeared 24 times above light pollution, of which 50% occurred in the first quarter and 41.7% occurred in the fourth quarter. In 2018, there were a total of 23 air quality index categories with light pollution or above. Among them, the number of occurrences in the first quarter accounted for approximately 43.5%, and the number of occurrences in the fourth quarter accounted for approximately 30.4%. The second main component explains the summer effect of air quality every year. In 2017 and 2018, the number of air quality index categories in
the second and third quarters was excellent and good were 255 and 249, respectively, and the proportion of occurrences was 60.4% and 56.6%.

According to the accumulation of the first two principal components of the 13 detection stations, the scatter plots corresponding to all the stations are placed. If the first principal component is positive and has a value, it means that the air quality of the monitoring station is poorer, on the contrary, the better the air quality. Similarly, the accumulation of the second principal component means that the air quality of the test station is worse in summer.

| Area     | PC1    | PC2    | PC3    | PC4    | Area     | PC1    | PC2    | PC3    | PC4    |
|----------|--------|--------|--------|--------|----------|--------|--------|--------|--------|
| Chengqu  | 191.82 | -23.92 | 22.06  | -7.11  | Jianghan | -164.83| -99.09 | -95.49 | -15.02 |
| Caidian  | 167.62 | 3.93   | -76.96 | 67.34  | Jiangxia | -215.51| -27.34 | 49.65  | -25.74 |
| Jiangan  | 84.15  | -61.61 | 34.89  | 47.54  | Qiaokou  | -120.18| 93.58  | -11.63 | 0.8    |
| Hannan   | -234.83| -130.8 | -48.5  | -14.17 | Qingshan | 357.21 | 63.29  | -1.45  | 7.38   |
| Hanyang  | 154.6  | -47.72 | 103.04 | 16.58  | Dongxihu | 277.52 | 14.62  | 3.87   | -86.78 |
| Hongshan | -185.32| -2.39  | -3.43  | -25.74 | Wuchang  | 183.03 | -43.57 | -45.53 | -7.69  |
| Huangpi  | -202.67| 220.48 | -7.63  | 4.99   | Xinzhou  | -292.61| -71.68 | 76.07  | 15.32  |

It can be seen from Figure 3 that the detection areas of the third quadrant (Xinzhou District, Huanan District, Jiang xia District and Jiang han District) have negative first and second principal component scores, indicating that the air quality index AQI in these regions is relatively low throughout the year is the area with the best air quality in Wuhan. In the detection area of the first quadrant (Cai dian District, Dong xi hu and Qing shan District), the first and second principal component scores are positive, and the first principal component score is higher, indicating that these areas are most affected by the winter effect. The region with the worst air quality in the city. In the detection area of the second quadrant (Huang bei District, Hong shan District and Qiao kou District), the score of the first principal component is negative, and the score of the second principal component is positive, which means that the air quality changes in these areas are more obvious in summer. For the fourth quadrant detection area (Hanyang District, Wuchang District, Jiang an District and the urban area representing the average level), the air quality is close to the average level of air quality in Wuhan. Therefore, the functional principal component analysis of the air quality curve of Wuhan City has explained the spatiotemporal changes of air quality in Wuhan to a certain extent, and the principal component score map can be used to distinguish 13 regions according to the different conditions of air quality. Divided into four groups: Cai dian District, Dong xi hu and Qing shan District as the first group; Huang pi District, Hong shan District and Qiao kou District as the second group; Xin zhou District, Huanan District, Jiang xia District and Jiang han District as the third group Group; Han yang District, Wuchang District and Jiang an District are the fourth group. In summary, it shows that the air quality in Wuhan has significant seasonal effects and regional characteristics.

3.2.3. Spatial distribution characteristics of air quality in Wuhan

The upper level divided the 13 areas in Wuhan into 4 categories according to the similarity of air quality changes according to the main component of each monitoring station. In order to reflect the changing trend of air quality in various regions, the ranking of air quality in 2017 and 2018 was calculated here, resulting in a visual display on the map, and further analysis of the spatial distribution characteristics of air quality. The results are as follows:
Table 2 Ranking of Wuhan by Region

| Area    | 2017 | 2018 | Area    | 2017 | 2018 |
|---------|------|------|---------|------|------|
| Xinzhou | 1    | 1    | Hanyang | 8    | 7    |
| Jiangxia| 2    | 2    | Hongshan| 9    | 5    |
| Hannan  | 3    | 3    | Wuchang | 10   | 12   |
| Huangpi | 4    | 4    | Chengqu | 11   | 11   |
| Jianghan| 5    | 7    | Caidian | 12   | 10   |
| Qiaokou | 6    | 6    | Donxihu | 13   | 13   |
| Jiangan | 7    | 9    | Qingshan| 14   | 14   |

As shown in Table 2, the air quality in Xinzhou District, Jiangxia District, Hannan District and Huangpi District has been at a relatively good level. Compared with other regions, the air quality in Dongxihu and Qingshan Districts is not optimistic. The air quality rankings of Jianghan District, Wuchang District and Jiang'an District have decreased compared to 2017. The air quality rankings of Hanyang District, Hongshan District and Caidian District have improved.

Figure 3 Spatial distribution characteristics of air quality in Wuhan in 2017 and 2018

The reasons for the relatively poor air quality in some areas of Wuhan City are mainly due to the following points: First, considering the wind direction factors, Wuhan is mainly based on the northeast wind direction, and the southwest wind direction is supplemented. Taking Qingshan District as an example, Qingshan District is not only in the downwind area but also has the East Lake Scenic Area and Huashan Ecology in the southwest as an environmental isolation zone, making Qingshan District the best place for heavy industry site selection, of which Qingshan District has Wuhan City Fire There are 4 power plants, 1 waste incineration power plant, 1 thermal power plant in Caidian District, 1 each in Dongxihu District Thermal Power Plant and 1 waste incineration power plant. The energy consumption of these heavy industries has made these prefecture-level cities a certain extent. The areas where the air quality in Wuhan is most in need of improvement.

4. The empirical results

Wuhan City is located in the middle and lower reaches of the Yangtze River Plain and the north of the Jianghan Plain. The terrain is mainly assisted by the main hills of the plain. There are a large number of lakes in the city. The water area accounts for about 25% of the city’s total area. The terrain is high in the east and low in the west. In terms of distribution, the air quality of Wuhan City in summer and autumn is better, mainly because of the abundant rainfall in summer and autumn, Wuhan has a certain supplementary effect on the pollutants in the air, and the air pressure in summer and autumn is relatively small. To facilitate air flow and produce a certain dilution and diffusion of pollutants in the air. For the spring and winter seasons, Wuhan City is the prefecture-level city with the best air quality in the slightly higher area in terms of spatial distribution. For Huangpi District and Xinzhou District, although they will be affected by the air in the surrounding areas to a certain extent, they have a wide area, large environmental capacity and open surrounding areas, few heavy industry enterprises and good air circulation, so their air quality is excellent; For Hannan District, most of the time is in the upwind
direction, and it is less affected by the surrounding areas; for Jiangxia District, it has a unique geographical advantage, even relatively large. There are a large number of lakes and wetlands, and in the northeast direction, there are East Lake Scenic Area and The Huashan Ecology serves as an environmental isolation zone and serves as a natural barrier so that it has always had good air quality.

Acknowledgments
This work is partially supported by Hubei Province Key Laboratory of Systems Science in Metallurgical Process (Wuhan University of Science and Technology) with Grant No. Z201901

References
[1] Cheng, Zhonghua. The spatial correlation and interaction between manufacturing agglomeration and environmental pollution[J]. Ecological Indicators, 2016, 61:1024-1032.
[2] Li K, Lin B. Economic growth model, structural transformation, and green productivity in China[J]. Applied Energy, 2017, 187(Complete):489-500.
[3] Spatial and temporal variability of ultrafine particles, NO2, PM2.5, PM2.5 absorbance, PM10 and PMcoarse in Swiss study areas[J]. Atmospheric Environment, 2015, 111:60-70.
[4] Cheam, A. S. M.M. Marbac, and P. D. Mcnicholas. "Model-based clustering for spatiotemporal data on air quality monitoring." Environmetrics 28.3(2017):e2437.
[5] Junshan W, Guojie S(2018). A Deep Spatial-Temporal Ensemble Model for Air Quality Prediction[J]. Neurocomputing, 314(1):198-206.
[6] Tian Y, Jiang Y, Liu Q, et al. Temporal and spatial trends in air quality in Beijing[J]. Landscape & Urban Planning, 2019, 185:35-43.
[7] Gang X, Limin J, Suli Z, et al. Examining the Impacts of Land Use on Air Quality from a Spatio-Temporal Perspective in Wuhan, China[J]. Atmosphere, 2016, 7(5):62-70.
[8] Silverman B W. Function estimation and functional data analysis[M]/First European Congress of Mathematics Paris, July 6–10, 1992. 1994.
[9] Leng X, Muller, H.-G. Classification using functional data analysis for temporal gene expression data[J]. Bioinformatics, 2006, 22(1):68-76.
[10] Auton T. Applied Functional Data Analysis: Methods and Case Studies[J]. Journal of the Royal Statistical Society, 2010, 167(2):378-379.
[11] Jacques J, Preda C. Model-based clustering for multivariate functional data[J]. Computational Statistics & Data Analysis, 2014, 71(3):92-106.
[12] Tzeng S L, Hennig C, Li Y F, et al. Dissimilarity for functional data clustering based on smoothing parameter commutation[J]. Statistical Methods in Medical Research, 2017:096228021771005.