Functional data analysis of air pollution in six major cities

Weiwei Xiao and Yan Hu1
College of Sciences, North China University of Technology, Beijing 100144, China
1Email: 1017480982@qq.com

Abstract. With the development of the economy, the problem of air pollution has become increasingly serious. This can be seen from the growth of haze in recent years. The main components of “fog and haze” are sulfur dioxide, nitrogen oxides, and particulate matters. And we call the particulate matters with diameters less than or equal to 2.5 micrometers PM2.5. This paper uses the PM2.5 indexes of the six major cities, Zhengzhou, Tangshan, Jinan, Nanjing, Wuhan and Beijing as parameters in functional data analysis to determine its relationship with PM10, O3 and temperature. As a result, we found out the interactive and restrictive relationships of these four factors. Finally, we propose some suggestions toward the treatment of air pollution.

1. Introduction

1.1. Research background and its significance
With the development of economy, the problem of air pollution has become increasingly serious. If the problem of environmental deterioration caused by economic development is not dealt with in time, then it will in turn constrain the development of the economy and at the same time greatly affect people's quality of life. The governance of air pollution, whether it is from China's economy, production or people's lives, is an imminent event.

1.2. The innovation of this article and its contribution
At present, functional data analysis is rarely applied in environmental metrology. In this paper, functional data analysis is applied to the study of the relationship between atmospheric pollutants PM2.5 and atmospheric PM10, O3 and temperature. The relationship between PM2.5 and PM10, O3 and temperature is determined from the perspective of the function. Furthermore, it can provide suggestions for the prevention and treatment of atmospheric pollution. The functional data model makes full use of the historical data, without any subjective judgment.

2. The index of air pollution and its research methods

2.1. some concepts
PM, the full name in English, is particulate matter. PM2.5 refers to particles in the atmosphere that are less than or equal to 2.5 microns in diameter. It is also called particulate matter that can enter the lungs. For PM10 it refers to solid and liquid particles with aerodynamic diameters below 10 microns. Their units are \( \mu g \cdot m^{-3} \) (micrograms per cubic meter). The larger the value, the more serious the air pollution. O3, also known as superoxide, is an allotrope of oxygen. At ordinary temperature, it is a light blue gas with a special odor. Ozone concentration refers to the amount of ozone per unit volume. Using mass
ratio units $mg/L$, $g/m^3$. The unit of temperature is in degrees Celsius and is represented by the symbol “°C”.

2.2. Functional data analysis
This paper uses the basic function smoothing method, the basic function smoothing method can convert discrete observations into corresponding functions, and the basic functions can retain the properties of the function to be estimated. The function $x(t)$ is estimated by the linear combination of basic functions $\phi_k(t) = (k = 1, L, K)$, which is expressed as $x(t) = \sum_{i=1}^{\hat{\lambda}} c_i \phi_i(t)$.

In the above equation, the smoothness of the observation data $y_j$ is determined by the number of basic functions. The more the number of basic functions, the less smooth the fitted function, but the higher the degree of fitting is. Given the corresponding basic function $\phi_i(t)$, the unique function is determined by the coefficient vector $c = (c_i, L, c_i)$ . When estimating the coefficient vector $c = (c_i, L, c_i)$, the least-squares method is usually used to minimize the sum of squared residuals

$$SMSE(y | c) = \sum_{j=1}^{n} \left[ y_j - \sum_{k=1}^{\hat{\lambda}} c_k \phi_k(t_j) \right]^2$$

It can also be written in the following matrix form $SMSE(y | c) = (y - \phi c)^T(y - \phi c) = \|y - \phi c\|^2$

The final estimated coefficient $c$ solution is $c = (\phi^T \phi)^{-1} \phi^T y$.

The basic functions are further divided into Fourier bases, polynomial bases, polynomial spline bases, and wavelet bases. In this paper, the Fourier basis function is used. Its expression is

$$x(t) = c_0 + c_1 \sin(wt) + c_1 \cos(wt) + c_2 \sin(2wt) + c_3 \cos(2wt) + \ldots$$

In this paper, we use functional regression analysis model in functional data analysis. The regression model is $y(t) = X^T \beta(t) + \varepsilon(t)$ $\varepsilon(t)$ is the error term. ([1, 2, 3])

3. Air pollution analysis

3.1. Data Sources
The research data in this paper comes from the National Environmental Protection Bureau data center and the Data sharing center of the National Meteorological Administration. This article contains daily PM2.5, PM10, O_3, temperature data in 2017 in Zhengzhou, Wuhan, Tangshan, Nanjing, Jinan and Beijing.

3.2. Data Fitting

3.2.1. Determination of penalty factors. Penalty factor calculation formula ([11] P96-97)

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

We want to find a smaller GCV as a penalty factor for the curve, so that smoothing the curve pays a smaller residual cost. Since the value of INDEX does not have a significant effect on the fitting result, INDEX can reduce the amount of the next operation step by taking an integer, so we might as well calculate integers for INDEX.
From figure 1, we can conclude that the GCV curve is a monotonically increasing curve. So take INDEX=1. The value is taken as a rough penalty factor to the next operation step.

3.2.2. PM2.5, PM10, $O_3$, temperature curve of six major cities. According to penalty factor, functional data fitting was performed for PM2.5, PM10, $O_3$ and temperature in six cities, Zhengzhou, Wuhan, Tangshan, Nanjing, Jinan and Beijing. In figure 2-8, the abscissa indicates each day, from Jan. 1 to Dec. 31 in 2017.

Figure 1. GCV Graph.

Figure 2. PM2.5 fittings of six cities.
Figure 3. PM10 fittings of six cities.

Figure 4. O3 fittings of six cities.
In Figure 2, the fitting line shows the trend of PM2.5. From the analysis of time, it can be seen that there are two obvious pulse points in the six cities in November-December and January-February. This shows that the peak value of atmospheric pollutants in China has a more obvious pattern. The cold weather in winter requires energy to burn and heat, making the air produce more particulate matter.

In Figure 3, compared with the trend analysis of PM10 in the six major cities, the more obvious three pulse points are November-December and January-February, April-May, of which Beijing, Tangshan, Zhengzhou, and Jinan are similar, and Tangshan and Beijing have basically the same trend. The PM10s in Zhengzhou and Jinan have similar degrees, and Wuhan has a greater similarity with the trend of PM10 in Nanjing.

In Figure 4, from the fitted line, we can see that the pulse time points of the six cities are almost the same in summer time. High summer temperatures have a significant effect on ozone levels.

3.2.3. Analysis of the relationship between PM2.5 and PM10, $O_3$, Temperature. In order to understand PM2.5, the major atmospheric pollutant, we used a functional regression analysis to fit PM2.5 and PM10 in six major cities. Among them, PM2.5 data has been functionalized. PM10 as a scalar. The univariate scalar corresponding variable model with intercepts was selected to fit the data of six cities and the relationship between PM2.5 and PM10 was obtained. As shown in Figure 6. (The abscissa is time t ([1:365]), and the ordinate is the fitting coefficient. It is the change of PM2.5 to PM10 as time t changes.)
From the trend of similar sinusoidal functions presented by the coefficient series in the above figure, it can be seen that the effect of PM2.5 on PM10 is different in different seasons. We know that PM10 refers to solid and liquid particles with aerodynamic diameters below 10 microns. PM2.5 refers to particles in the atmosphere that are less than or equal to 2.5 microns in diameter. By definition, we know that particles belong to PM2.5 also belong to PM10, but through the analysis of them, PM2.5 was negatively correlated with PM10 in the early spring and autumn and winter. PM2.5 is positively correlated with PM10 in spring and summer, which is consistent with the increase of suspended particles in the air in summer.

This article also studied the relationship between O₃ and PM2.5. Figure 7 shows the relationship between PM2.5 and O₃.

From the trend of the PM2.5 and O₃ coefficient series in the above figure, we find that it is strikingly similar to the sequence of PM2.5 and PM10 coefficients, and that PM2.5 is negatively related to O₃ in the early spring and autumn and winter. PM2.5 is positively related to O₃ in spring and summer, and it can be seen from the figure that the amplitude gradually decreases in one year. This shows that the impact between O₃ and PM2.5 is smaller in autumn and winter.

In figure 8 the relationship between temperature and PM2.5 is shown.
In figure 8 we can see that the influence of temperature on PM2.5 are different. In spring and summer, PM2.5 and temperature show a negative correlation. But in the autumn and winter time PM2.5 and The temperature is positively correlated. This is related to the fact that the temperature of autumn and winter is low and the flow velocity of air is slower, which makes PM2.5 less prone to spread.

PM2.5 is the "culprit" of the formation of smog. The indicators that influence it are not only PM10 and O3 but also temperature and atmospheric pollution is the result of a combination of different factors. In order to fundamentally govern atmospheric pollution, it is necessary to further understand PM2.5, the main factor in atmospheric pollution.

4. Suggestions for the treatment of air pollution

Based on the above analysis, we can conclude that PM2.5 and PM10 and O3 and temperature are interrelate and restrict mutually. How we should prevent and control it is now the top priority and implementation.

4.1. Air pollution control measures

Figure 6 and figure 7 show that PM10 and O3 are positively correlated with PM2.5 in spring and summer, so at this time, joint prevention and control of O3 pollution must be done. At the same time, in order to reduce the content of PM10 and PM2.5 we also need to take corresponding measures. This will be followed by some specific recommendations. From the relationship between PM2.5 and temperature in figure 8, it can be concluded that the increase in PM2.5 not only harms people's health, but also slowly changes our climate. Therefore, the management of PM2.5 is imminent. The following gives some advice on the governance of PM2.5 in terms of industry and transportation.

In industrial field, we should strictly control high energy consumption, increase production capacity in high pollution industry, eliminate backward technology and equipment, actively adopt clean production technology and comprehensive utilization technology in related industries, and vigorously cultivate environmental protection and new energy industry. For projects that fail to meet environmental targets, no water supply or electricity will be allowed.

In traffic field, encourage people to travel by public transport, encourage the development of bicycle transportation, strengthen the environmental protection supervision of vehicles, and solve the car exhaust gas by installing automotive catalytic converters. The fuel design and fuel quality of the engine are changed, and traffic management is strengthened. Expand the transport scope and capability of the MTR, promote container rail and rail transport, and speed up the proportion of railway transportation. Accelerate the elimination of old motor vehicles and eliminate the emission of pollutants from old cars.

For the governance of air pollution, joint prevention and control will be an inevitable trend. The establishment of the Bohai Rim, including the Beijing-Tianjin-Hebei region, the Yangtze River Delta,
the Pearl River Delta and other regions, will strengthen joint prevention and control, strengthen PM2.5 governance in densely populated areas and major cities, and build Responsibility assessment system for atmospheric environment improvement. Establish a code of conduct for the whole society to "live together and fight together”.

5. Conclusion remarks
The paper fits the data of PM2.5 through fitting function data of PM2.5, PM10 and O₃ and temperature indicators of the six cities, Zhengzhou, Wuhan, Tangshan, Nanjing, Jinan and Beijing in 2017. The relationship between PM2.5 and PM10 and O₃ and temperature was studied using a Fourier-based smoothing method and then a functional regression method. We can conclude that PM2.5 and PM10 and O₃ and temperature are interrelated and restrict mutually. In addition, measures and suggestions for the prevention and control of atmospheric pollution.

Acknowledgments
This paper is supported by Beijing Natural Science Foundation (1152001). The authors thank the anonymous reviewers for the valuable comments and suggestions to the paper.

References
[1] Ramsay J O and Silverman B W 2005 Functional Data Analysis Second Edition
[2] Ramsay J O and Gils Hooker Spencer Graves Functional Data Analysis with R and MATLAB
[3] Ramsay J O 1996 Estimating smooth monotone functions Journal of the Royal Statistical Society series B60 365-375