Research on Data Calibration of Air Quality

Qiming Feng¹,*

¹Department of Mathematics, Wuxi Institute of Commerce, Wuxi, Jiangsu, 214135, China
*e-mail: fengqiming@wxic.edu.cn

Abstract. In recent years, the urban air pollution in our country has become more and more serious, which has aroused widespread concern of the general public and the scientific community. The micro air quality detector not only costs little, but also can real-time monitor the air quality of a certain area in a grid way, so it can be used as the supplement of national survey point data. Based on the canonical correlation analysis of the data, it is found that the concentration deviation of "two dust and four gas" is significantly related to the meteorological parameters, among which the concentration deviation of PM2.5, PM10, NO2 and O3 is greatly related to the factors of pressure and humidity, and it is also known that the correlation between concentration deviation and humidity is the largest. And the concentration deviation between self-built point and national survey point is modeled. The results of this study can provide a method for the completion of urban air quality data, and the research method can provide a reference for data mining.

1. Introduction

In recent years, air pollution has been increasing in China's cities, and the main is haze pollution, which has aroused widespread concern of the general public and the scientific community. Through the real-time monitoring of the concentration of "two dusts and four gases" (PM2.5, PM10, CO, NO2, SO2, O3), the air quality can be grasped in time and the corresponding measures can be taken for the pollution source. [1] Although the national monitoring and control station (national survey point) has monitoring data on "two dusts and four gases", which is more accurate. However, due to the small number of national survey points, the long-time lag of data release, and the large cost, it is unable to provide real-time air quality monitoring and prediction.

In order to solve these problems, a company has independently developed a micro air quality detector, which not only has a small cost, but also can real-time monitor the air quality of a certain area in a grid way, and can simultaneously monitor temperature, humidity, wind speed, air pressure, precipitation and other meteorological parameters. Because of these factors, such as a certain zero drift and range drift for the chemical gas sensor being used for a long time, cross interference to the sensor due to the change of unconventional pollutant concentration, and the influence of weather factors on the sensor, there is a certain difference between the data collected by the micro air quality detector and the data of the national survey point at the same time. Therefore, it is necessary to use the hourly data of national survey point to calibrate the data of self-built point near the national survey point.

Here, based on the data of question C of national mathematical model competition for college students in 2019, the following three topics are mainly studied. [2]

(1) The data of self- -built point and national survey point will be analysed.
(2) The factors that cause the difference between self-built point data and national survey point data will be analysed.

(3) Using the data of national survey point, the mathematical model of self-built point data calibration will be established.

This paper analyses the given data through data exploratory and multiple regression, finds out the abnormal data in time, removes the suspicious or unreasonable data, and analyses the causes of the abnormal, so that the objective environmental quality can be truly reflected.

2. Data exploratory

For the data collected by the micro air quality detector and the national survey point, exploratory analysis is carried out to observe the missing data and abnormal data, so as to improve the accuracy and effectiveness of air quality data analysis. Using the data, the curve of the concentration of "two dust and four gas" with time can be made.

There are 234717 and 4200 data collected from self-built points and national survey points respectively, and the time intervals are more than five minutes and one hour respectively. It can be seen from the above that there is a lack of data in the data collected from self-built points and national survey points, and there are some high or low points in the graph, which belong to data outliers. For future model, it is necessary to establish standards to eliminate outliers. The following only gives the curves of $PM2.5$ and $CO$ in the data collected from self-built points and the national survey points. Where, the horizontal axis represents time, and the vertical axis represents concentration.

Figure 1. The curves of PM2.5 and CO in the data collected by the national survey points
3. Data difference

In this paper, the data differences and meteorological parameters are analysed to find out the influencing factors. For two groups of variables \( x_1, x_2, \ldots, x_m \) and \( y_1, y_2, \ldots, y_n \), the so-called canonical correlation analysis is to study the correlation between two groups of variables. Two variables \( U \) and \( V \), which represent the linear combination of variables in two groups, are selected to form a representative comprehensive index. By studying the correlation between the two groups of comprehensive indexes, the overall correlation between the two groups is reflected.

For the sake of simplicity, the average value of the concentration of "two dust and four gases" from the self-built points is represented by \( Z_i \), \( G_i \) is the concentration from the national survey points, and \( \Delta P_i \) is the concentration deviation, that is,

\[
\Delta P_i = Z_i - G_i.
\]

In order to analyse the factors that cause the difference between self-built point data and national survey point data, a set of variables \( x_1, x_2, \ldots, x_6 \) denotes \( PM2.5, PM10, CO, NO^2, SO_2, O_3 \) respectively, another set of variables \( y_1, y_2, \ldots, y_5 \) means wind speed, air pressure, precipitation, temperature, humidity respectively.

According to the given data, the canonical correlation analysis of \( X \) and \( Y \) is carried out, that is,

\[
\begin{align*}
U &= 0.3475x_1 + 0.3456x_2 - 0.0554x_3 + 0.3094x_4 + 0.1105x_5 + 0.3767x_6 \\
V &= -0.1118y_1 + 0.4555y_2 + 0.2981y_3 + 0.1112y_4 + 0.7846y_5.
\end{align*}
\]

The canonical correlation coefficient is \( \rho = 0.7901 \), and the following figure shows the correlation of \( U_i \) and \( V_i \) variables.
Figure 3. The scatter diagram of canonical related variables

Finally, the canonical correlation coefficient is tested ($\alpha = 0.05$), and the test results are shown in table 1.

| $k$ | $A_k$ | $F_k$   | $d_{1k}$ | $d_{2k}$ | $P_k$ |
|----|-------|---------|----------|----------|-------|
| 1  | 0.2225| 245.7161| 30       | 16166    | 0     |
| 2  | 0.5923| 114.6684| 20       | 13407    | 0     |
| 3  | 0.7827| 86.5001 | 12       | 10697    | 0     |
| 4  | 0.8978| 74.663  | 6        | 8088     | 0     |
| 5  | 0.9909| 18.4864 | 2        | 4045     | 0     |

Because $P_k$ is equal to 0, each pair of canonical variables are significantly related. Since $U$ mainly reflects the concentration deviation information of $PM_{2.5}, PM_{10}, NO_x, O_3$, and $V$ mainly reflects the pressure and humidity, the pair of canonical variables mainly reflect the concentration deviation of $PM_{2.5}, PM_{10}, NO_x, O_3$, which is highly related to the pressure and humidity, especially the humidity.

4. Data calibration

4.1. Model establishment

For the calibration of air quality data, there have been some researches at home and abroad. At present, the addition correction and multiplication correction are two widely used methods in data processing. Assuming that the data of national survey points are accurate, the difference between the data of self-built points and the data of national survey points is made to find the law of deviation distribution. In this paper, the addition correction method is used to calibrate the data of self-built points. According to the given data, the concentration deviation can be fitted as a function of time $t$, that is,

$$\Delta P_i(t) = a_i t^2 + b_i t + c_i$$

Where $i = 1, 2, \cdots, 6$ respectively indicate that the corresponding substances are $PM_{2.5}, PM_{10}, CO, NO_x, SO_2$, and $O_3$. The coefficients of the fitting function are as follows.

| $i$ | $a_i$ | $b_i$   | $c_i$    |
|-----|-------|---------|----------|
| 1   | -0.0008| 0.0730  | 19.8755  |
| 2   | -0.0009| -0.2680 | 75.3294  |
| 3   | 0.0001 | -0.0112 | -0.1151  |
| 4   | 0.0036 | -0.9541 | 71.0668  |
| 5   | -0.0013| 0.3487  | -18.9068 |
| 6   | 0.0016 | -0.08015| 71.5282  |

In this way, the calibration model of the data collected from the self-built points can be obtained as follows.
\[ J_i(t) = Z_i \Delta P_i(t). \]

4.2. Model check

According to the data of national survey points, the error analysis of the data calibrated is carried out. Here, the mean absolute error and mean relative error are selected as the evaluation indexes. In the given data, we randomly select three groups of data, including 1000, 10000 and 100000 data respectively, the calculated mean absolute error (MAE) and mean relative error (MRE) are shown in table 3.

Table 3. The mean absolute error and mean relative error

| Number   | PM2.5 | PM10 | CO  | NO2 | SO2 | O3  |
|----------|-------|------|-----|-----|-----|-----|
|          | MAE   | MRE  | MAE | MRE | MAE | MRE |
| 1000     | 16.5  | 28.9 | 22.3| 31.1| 0.1 | 12.0|
| 10000    | 16.6  | 28.5 | 23.1| 31.7| 0.1 | 12.1|
| 100000   | 16.5  | 28.4 | 23.1| 31.6| 0.1 | 12.1|

It can be seen from table 3 that the mean absolute error and the mean relative error are relatively stable, which shows that the model established in this paper has a good effect on data calibration.

5. Conclusion

Based on the canonical correlation analysis of the data, it is found that the concentration deviation of "two dust and four gas" is significantly related to the meteorological parameters, among which the concentration deviation of PM2.5, PM10, NO2 and O3 is greatly related to the two factors of pressure and humidity, and it is also known that the correlation between concentration deviation and humidity is the largest. Therefore, when improving the detector, the monitoring data should not be affected by humidity as much as possible. In this paper, the concentration deviation of "two dust and four gas" between self-built point and national survey point is modeled. After the data of self-built point is calibrated, the mean absolute error and mean relative error of the model are calculated. The test results show that the model is stable and the calibrated data can be used to represent real-time air quality.

References

[1] Liao Zhiheng. (2015) Characteristics and Influencing Factors of Air Pollution in the Pearl River Delta from 2006 to 2013. China Environmental Science. 2015(2):329-336.
[2] 2019 question C. National Mathematical Model Competition for College Students. http://www.mcm.edu.cn/html_cn/node/a1ffce4e5587c8a6f96eacef8dbee34e.html.
[3] Yu Li, Song Lei. (2015) The Calibration of TRMM 3b43 Precipitation Data in the Qinghai Tibet Plateau from 1998 to 2012. Arid Geography. 38 (5): 900-911.
[4] Zhang Zhaonian, Zhang Caixiang. (2006) Analysis on the Comparability of Monitoring Ambient Air Quality Data by Passive Sampler. China Environmental Monitoring. 22 (4): 42-44.
[5] Zhang Beina, Feng Zhenhua, Zhang Feng, Du Zhenhong, Liu Renyi, Zhou Qin. (2019) Research on the Method of Urban Air Quality Data Completion based on Spatiotemporal Multi View BP Neural Network. Journal of Zhejiang University (Science Edition).46(6):737-774.
[6] Lei Bao. (2017) Exploratory Air Quality Data Analysis based on Big Data. Electronics World. 2017(6):122-130.
[7] Wang Shen, Cai Xue, Guo qiucen. (2016) Abnormal Value Monitoring of Air Quality Data. China New Telecommunications.18(18):79-86.