Beijing-Tianjin-Hebei Region Air Pollution Analysis

Weichen Yang

School of Electrical and Electronic Engineering, North China Electric Power University, Baoding, Hebei 071003, China.

949830657@qq.com

Keywords: Principal component analysis; clustering analysis; partial least squares regression; multiple regressions.

Abstract. Aiming at air pollution in Beijing-Tianjin-Hebei region, this paper carries out modeling analysis, considering different season pollutant composition and regional distribution differences; a kind of effective approach that analyzed air pollution relationship in the region is determined. In order to simplify the model, pollutant diffusion model is simplified based on the principal component analysis and clustering analysis, the overall diffusion trend of pollutant are obtained by multiple regression analysis. This method has good universality and the results are consistent with reality.

1. Introduction

In recent years, the air pollution situation becomes worse and worse in Beijing-Tianjin-Hebei region [1], the research of air pollution is more important in this region. This paper puts forward a kind of effective approach that analyzed regional air pollutants transfer relationship, which has a guiding significance for pollution prevention and control work in this region.

2. Analysis Method

2.1 Data Source.

The used data in this thesis derived from record data Chinese air quality monitoring analysis platform in 2015 [2].

2.2 Principal Component Analysis.

We frequently use PM2.5, PM10, SO_2, NO_x, O_3, CO six indexes to reflect air pollution degree. A certain relevance exists in six environmental indexes can be found through analysis, and the each component proportion of pollutants in different months is different. In order to get a comprehensive index that can reflect the pollution degree, the principal component analysis can be used, a linear combination of six pollution indexes get pollution degree comprehensive index Z, which is more comprehensive to reflect regional air quality. Taking the pollution situation of various cities in Beijing-Tianjin-Hebei district for example in January and analyzed, the cumulative contribution rate of the first two principal components was 88.878%, the expression formula are respectively:

$$y_1 = 0.439\tilde{x}_1 + 0.437\tilde{x}_2 + 0.394\tilde{x}_3 + 0.428\tilde{x}_4 - 0.300\tilde{x}_5 + 0.433\tilde{x}_6$$

$$y_2 = 0.089\tilde{x}_1 + 0.178\tilde{x}_2 + 0.226\tilde{x}_3 + 0.007\tilde{x}_4 - 0.939\tilde{x}_5 + 0.168\tilde{x}_6$$

Among them, the original air quality indexes take value after standardization. The corresponding comprehensive score function is:

$$Z = \sum_{j=1}^{p} b_j y_j = 0.878 y_1 + 0.122 y_2$$

$$= 0.396\tilde{x}_1 + 0.405\tilde{x}_2 + 0.374\tilde{x}_3 + 0.377\tilde{x}_4 - 0.378\tilde{x}_5 + 0.401\tilde{x}_6$$

Considering the different proportion of air pollutant components in different season, the data of each month are respectively needed to carry out principal component analysis.

2.3 Regional Division.

In order to simplify the model, clustering analysis can be used to classify the cities of the Beijing-Tianjin-Hebei region, thus reduce the operation amount. We believe that the city that air pollution situations are similar in a month and regions connected should be divided into one category, we used the sum of squares of deviations clustering method in the analysis. Taking for example the distribution situation of pollutants in January, The Beijing-Tianjin-Hebei region can be appropriately divided into four regions. Q_1 (Langfang, Beijing, Qinhuangdao, Tianjin, Cangzhou, Tangshan), Q_2 (Zhangjiakou, Chengde), Q_3 (Handan), Q_4 (Baoding, Xingtai, Shijiazhuang, Hengshui)

Considering that the distribution situation of air pollutants in the different season are different in the region, so when dividing regions need to be at the end of each month.

2.4 Regional Score Calculation.

After clustering analysis, the cities that air quality indexes are similar and degree of correlation are high are gathered for a category, form a new region. But the description of the new regional air quality indexes can't be simply average city comprehensive index in the region, also should consider the differences among the different city region. According to the size of cities, weighted average is carried out. Steps are as follows:

If divided new region $Q=\{q_1,q_2,\cdots,q_n\},(n\leq 13)$, then the regional comprehensive evaluation score is:

$$Z_Q=\frac{\sum_{i=1}^{n} z_i s_i}{\sum_{i=1}^{n} s_i}$$

s is region of the various cities. After weighted average, the regional comprehensive evaluation score can be gotten.

2.5 Regional Air Quality Influence Model.

As we know, there are several major sources in air pollutants at a certain moment in one region, respectively is: the day before yesterday the air pollutants (R), new generated pollutants and pollutants from outside of Beijing-Tianjin-Hebei region (N), spread from other regions (t). Air pollution also has several whereabouts: natural degradation (d), (t) spread to other regions. Taking interaction between the two regions for example, the pollutants transfer relations as Figure 1.

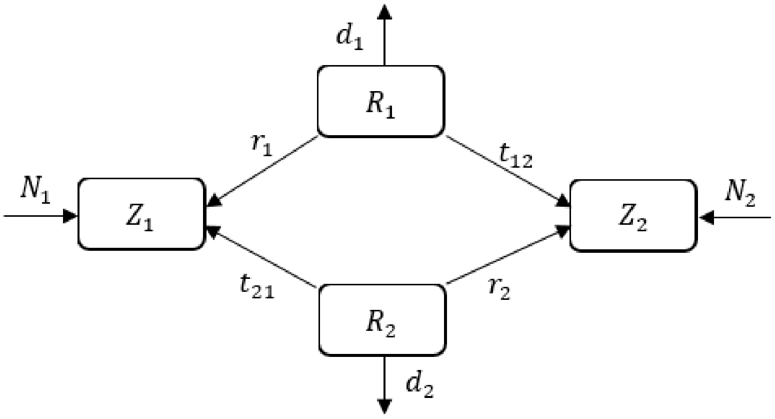

Figure.1 Pollutants transfer relations

r_i is represented pollutants retention rate of the i region.

Use t_{ij} represent ratio that pollutant transmit from the i region to the j region occupy total the day before.

$$Z_i=R_i\cdot r_i+\sum_{j=1}^{n}t_{ij}\cdot R_j+N_i$$

$$d_i + r_i + \sum_{j=1}^{n} t_{ij} = 1$$

Using the above two formula can complete modeling of regional air pollution mutual influence, it concisely and clearly reflects the mutual relationship among regions.

Partial further squares regression(PLSR:): it is a new kind of multiple statistical analysis methods, it mainly studies regression modeling that multiple dependent variables on multiple independent variables, especially when the internal height linearity of variables are relevant, partial further squares regression is more effective. In addition, partial further squares regression is better solve the problem such as sample number less than the number of variables [3]. According to the known data, we can carry out partial further squares regression to get r, t the two important indexes.

In the above analysis, we have got comprehensive scores in the all regions in the each month and each day. We will carry on data processing: taking January for example, the scores in various regions in January 1st as the independent variables, and the scores in various regions in January 2nd as the dependent variable, the scores in various regions in January 3rd as the dependent variable. By that analogy, the independent variable and dependent variable data set are obtained.

Use Matlab programming to realize partial least squares the mutual relationship among regions are obtained.

$$\begin{cases} Z_1 = 1.5136 + 0.2126R_1 + 0.9129R_2 - 0.2937R_3 + 0.0312R_4 \\ Z_2 = -0.9091 + 0.0871R_1 + 0.4460R_2 - 0.1697R_3 + 0.0006R_4 \\ Z_3 = 2.6218 + 0.1248R_1 + 1.2371R_2 - 0.2502R_3 + 0.3826R_4 \\ Z_4 = 3.8374 + 0.3224R_1 + 1.2079R_2 - 0.3244R_3 + 0.0772R_4 \end{cases}$$

Its transmission diagram is shown in Figure 2, and wind field diagram the month is shown in Figure 3.

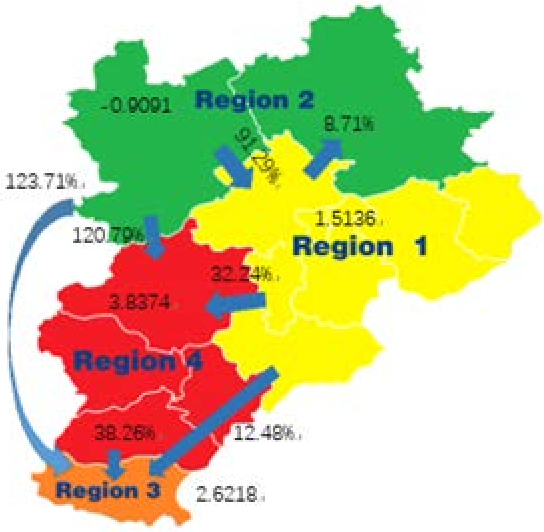

Figure.2 Transmission relation map

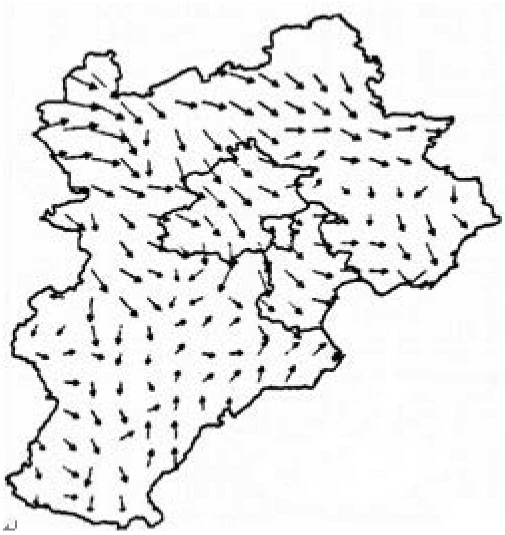

Figure.3 Wind field map

The average values in all region scores are respectively: -0.237876638, -1.785824591, 0.458271974, 1.584571751.

Thus we can see pollution severity degrees are respectively region 2, region 1, region 3, and region 4 from low to high. Constant term reflects the net generated new pollutants amount in various religions, it can be seen that region 4 is the main production source of the pollutants, and net production amount in the region 2 is minimum, it shows that the pollutant production in region 1 is the lowest in all regions, which are in conformity with the reality.

R_2, Z_3 and Z_4 regression coefficient is greater than 1, it seems not very reasonable, but after analysis, it is reliable. R_2 is corresponding to Zhangjiakou and Chengde two regions, because of the higher ground, and pollutants in other regions is not easy to transfer to the two cities, generally speaking, so pollutants can only transport from the two cities. The amount of pollution in the two regions is little; pollutants that transfer to other areas is more than local is entirely possible.

It can be found from Figure 2 and Figure 3, the pollutants of the Beijing-Tianjin-Hebei region on the whole spread from north to south, and it is consistent with the wind field Figure. The region 3 is most southern in the Beijing-Tianjin-Hebei region, which will be affected by other regions. When R3 increases, which mean more pollutants has diffused to region 3; it is naturally ease pollution condition of other regions. This also explains why R_3 and Z_1, Z_2, Z_4 are negatively correlated.

In the same analysis method, we also be able to get diffuse relationship other months, and not explained here.

3. Conclusion

Because when selecting different months to analyze, we must carry on principal component analysis for the index in that month, and the various cities are carried out clustering analysis, so it can more accurately reflects the pollutants diffuse relation among each month and has good universality.

In addition, due to a large amount of data analysis, reduce the influence of the special points on the whole and has a good stability.

Acknowledgement

If you follow the “checklist” your paper will conform to the requirements of the publisher and facilitate a problem-free publication process.

Many teachers in NCEPU in Baoding gave me much advice in this paper. I would like to express my gratitude to all those who helped me during the writing of this paper.

References

[1] P. Zhao, X. Zhang, X. Xu. (2011). Long-term visibility trends and characteristics in the region of Beijing, Tianjin and Hebei, China Atmos. Res., 101 (2011), pp. 711–718

[2] Chinese Air Quality Monitoring Analysis Platform, http://www.aqistudy.cn/

[3] Kramer, R. (1998). Chemometric Techniques for Quantitative Analysis. Marcel-Dekker. ISBN 0-8247-0198-4.