Air pollution analysis based on PCA and entropy weight method

Yingcan Chen*

Hajim School of Engineering and Applied Sciences, University of Rochester
Rochester, US

*Corresponding author: ychen192@u.rochester.edu

Abstract. With the rapid development of industrial engineering, more and more environmental problems have appeared. Atmospheric pollution has been a serious problem in modern society, and in this study, Principal Component Analysis (PCA) algorithm and Entropy weight method are employed to find the main pollutants in air. And Nanjing is applied as an example to analyze the pollutant, the result shows BaP and Acy is the biggest source of pollution.

Keywords: Atmospheric pollution; PCA; Entropy.

1. Introduction
PM2.5 is not only harmful to human health and the environment, but also an important cause of reduced visibility in urban atmosphere. Therefore, it has become a hot topic in the field of environmental research in recent years. PM2.5 in China has become the focus of many case studies. Atmospheric fine particulate matter contains solid and liquid aerosol particles, its composition is very complex, is the gas mixture of many substances. Particles with aerodynamic equivalent diameters of less than 2.5 μm in ambient air are called as Fine Particulate Matter PM2.5. Long-term exposure to PM2.5 will seriously affect the health of the body, in particular, it will induce the occurrence, development and deterioration of respiratory diseases. Therefore, in-depth research on the quality concentration of atmospheric PM2.5, the pollution characteristics of chemical components and the analysis of pollution sources in Nanjing provides a basis for Wuhan to solve the problem of atmospheric fine particulate pollution. It is extremely important for the management department to formulate reasonable measures and countermeasures, so as to improve the living environment and reduce the threat to human health. PCA and entropy methods are applied to find the source of pollution in air.

2. Application of PCA
PCA is a statistical method, which can be applied to matrix compression. While reducing the dimension of the matrix, the information of the original matrix can be retained as much as possible, and only the main characteristics existing in the matrix can be retained, thus greatly saving space and data amount [1]. In this paper, the process of implementing the algorithm to find the most important pollutant.
   
   Step 1: Determine the original indicators and conduct standardized processing
\[ X = \begin{pmatrix}
  x_{11} & x_{12} & \cdots & x_{1j} \\
  x_{21} & x_{22} & \cdots & x_{2j} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{i1} & x_{i2} & \cdots & x_{ij}
\end{pmatrix} \]  

(1)

\( x_i \) in the formula represents the \( i \) index of the \( j \) sample unit, and the data are standardized by the following formula (2):

\[ x_{ij}' = \frac{x_{ij} - \bar{x}_j}{s_j} \]  

(2)

Where \( \bar{x}_j = \frac{1}{m} \sum_{i=1}^{m} x_{ij}, s_j = \frac{1}{m-1} \sum_{i=1}^{m} (x_{ij} - \bar{x}_j)^2 \)

**Step 2:** Use SPSS software to achieve the PCA.

The data of Nanjing is input to model, which include 15 original pollutants, they are Nap, Acy, Flu, Ace, Phe, Ant, Fla, Pyr, Chr, BaA and BbF. PCA principal component dimension reduction toolbox in statistical software SPSS 12.0 was used to input standardized data matrix for dimension reduction processing [2]. The output results of SPSS are summarized in Table 1, Table 2 and Figure 1. It can be obtained from Table 1 that the cumulative variance contribution rate of the first four components reaches 91.202%. Meanwhile, it can be seen from the output gravel figure 1 that the distribution curve of characteristic roots gradually tends to be flat after the fourth characteristic root, so it is reasonable to extract the first three principal components. Table 2 is the matrix of principal component coefficients.

**Table 1. PCA result**

| Principal component | Variance | Cumulative contribution rate |
|---------------------|----------|-----------------------------|
| 1                   | 58.110   | 58.110                      |
| 2                   | 19.255   | 77.365                      |
| 3                   | 13.838   | 91.202                      |
| 4                   | 5.037    | 96.240                      |
| 5                   | 1.863    | 98.103                      |
| 6                   | 0.731    | 98.834                      |

**Fig.1** Scree plot of PCA
Table 2. Component coefficient matrix

| Principal component | 1     | 2     | 3     |
|---------------------|-------|-------|-------|
| 1                   | -0.208| 0.753 | -0.488|
| 2                   | 0.814 | -0.105| -0.229|
| 3                   | 0.127 | 0.891 | -0.018|
| 4                   | 0.099 | 0.97  | 0.1   |
| 5                   | 0.304 | 0.689 | 0.61  |
| 6                   | 0.43  | 0.291 | 0.778 |
| 7                   | 0.802 | -0.271| 0.457 |
| 8                   | 0.851 | -0.2  | 0.416 |
| 9                   | 0.904 | -0.156| 0.208 |
| 10                  | 0.942 | -0.133| 0.189 |
| 11                  | 0.955 | 0.029 | -0.236|
| 12                  | 0.961 | 0.011 | -0.246|
| 13                  | 0.964 | 0.033 | -0.202|
| 15                  | 0.927 | 0.05  | -0.27 |
| 16                  | 0.894 | 0.212 | -0.324|

3. Application of Entropy weight method

3.1. Data processing
The data of Nanjing is used herein. The first step is to normalize the data, which can guarantee the accuracy of prediction model and Z-scores method [3] is employed herein, which can be expressed as Equation (3):

\[ z = \frac{x - \mu}{\sigma} \]  

Where \( x \) is the original data;  
\( \mu \) is the mean value of the original data;  
\( \sigma \) is the variance of the original data.

3.2. Entropy weight method
Entropy weight method is an objective weighting method, which determines the weight of each index according to the information provided by the observed value of each index. In information theory, entropy is a measure of uncertainty. The more information there is, the less uncertainty there is, and the lower entropy there is; The smaller the amount of information, the greater the uncertainty and the greater the entropy. According to the characteristics of entropy, we can judge the degree of randomness and disorder of a scheme by calculating the entropy value, or judge the degree of dispersion of an index by using the entropy value. The greater the degree of dispersion of an index, the greater the influence of the index on the comprehensive evaluation. Therefore, the weight of each index can be calculated by using the tool of information entropy according to the variation degree of each index, and the process can be summarized as below:

**Step 1: Data matrix**

\[ A = \begin{pmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nm} \end{pmatrix}_{n \times m} \]  

**Step 2: Non-negative processing of data**
Entropy method is used to calculate the ratio of a certain index in each scheme to the sum of the same index value. Therefore, there is no dimensional influence, and there is no need for standardized processing. If there is a negative number in the data, the data should be de-negated! In addition, in order to avoid the meaninglessness of logarithms in entropy calculation, data translation should be carried out, and the larger the better the index [4]:

$$X_{ij}^\prime = \frac{x_{ij}^\prime - \min(x_{1j}, x_{2j}, \ldots, x_{nj})}{\max(x_{1j}, x_{2j}, \ldots, x_{nj}) - \min(x_{1j}, x_{2j}, \ldots, x_{nj})}$$  \hspace{1cm} (5)

For the indicators that less value indicates better performance:

$$X_{ij}^\prime = \frac{\max(x_{1j}, x_{2j}, \ldots, x_{nj}) - x_{ij}}{\max(x_{1j}, x_{2j}, \ldots, x_{nj}) - \min(x_{1j}, x_{2j}, \ldots, x_{nj})}$$  \hspace{1cm} (6)

For the sake of convenience, the data after the non-negative processing are still recorded to calculate the proportion of the $i$ scheme in the $j$ index [5]

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \quad (j = 1, 2, \ldots, m)$$  \hspace{1cm} (7)

**Step 3: Calculation of entropy**

$$e_j = -k \cdot \sum_{i=1}^{m} P_{ij} \log(P_{ij})$$  \hspace{1cm} (8)

$$k > 0, \quad e_j \geq 0$$

$$k = \frac{1}{\ln m}, \quad 0 \leq e \leq 1$$

**Step 4: Calculation of Coefficient of difference**

$$g_j = 1 - e_j$$  \hspace{1cm} (9)

Note that bigger value of $g_j$ indicates more important.

**Step 5: Calculation of weight**

$$W_j = \frac{g_j}{\sum_{j=1}^{m} g_j}, \quad j = 1, 2, \ldots, m$$  \hspace{1cm} (10)

**Step 5: Calculation of comprehensive score**

$$S_i = \sum_{j=1}^{m} W_j \cdot P_{ij} \quad (i = 1, 2, \ldots, n)$$  \hspace{1cm} (11)

### 3.3. Application

MATLAB software is used to solve the weight based on entropy method, and the result is shown in Table 3 and Figure 2.
Table 3. Component coefficient matrix

| Variables | weight |
|-----------|--------|
| NaP       | 0.052851 |
| Acy       | 0.084178 |
| Flu       | 0.051648 |
| Ace       | 0.04441 |
| Phe       | 0.034329 |
| Ant       | 0.038401 |
| Fla       | 0.065482 |
| Pyr       | 0.063537 |
| Chr       | 0.066205 |
| BaA       | 0.082319 |
| BbF       | 0.069779 |
| BkF       | 0.073981 |
| BaP       | 0.087039 |
| DBA       | 0.062334 |
| IcdP      | 0.059001 |

Fig. 2 Entropy weight

Then the result of PCA and entropy weight method are summarized in Table 4.

Table 4 Component coefficient matrix

| PCA pollutant | proportion | Entropy pollutant | proportion |
|---------------|------------|-------------------|------------|
| Vehicle       | 50.23%     | Coke oven         | 8.42%      |
| exhaust       | 50.23%     | volatile          |            |
| emission      |            |                   |            |
| Biomass       | 21.68%     | Biomass           | 22.16%     |
| combustion    |            | combustion        |            |
| Coal          | 19.29%     | Coal              | 27.75%     |
| combustion    |            | combustion        |            |
| /             | /          | Vehicle           | 41.66%     |
| /             | /          | exhaust           |            |
| /             | /          | emission          |            |
4. Conclusion
The conclusions can be drawn that the main pollutants of air in Nanjing is vehicle exhaust emission, and coal combustion and biomass combustion are also a serious problem.

References
[1] Yang, J. , Zhang, D. D. , Frangi, A. F. , & Yang, J. Y. . (2004). Two-dimensional pca: a new approach to appearance-based face representation and recognition. IEEE Trans Pattern Anal Mach Intell, 26 (1), 131-137.
[2] Yang, W. , & Omaye, S. T. . (2009). Air pollutants, oxidative stress and human health. Mutation Research, 674 (1-2), 45-54.
[3] Coifman, R. R. , & Wickerhauser, M. V. . (1992). Entropy-based algorithms for best basis selection. IEEE Transactions on Information Theory, 38(2), 713-718.
[4] Shannon, & C. , E. . (2014). Prediction and entropy of printed english. Bell System Technical Journal, 30(1), 50-64.
[5] Dong, X. , & Lewkowycz, A. . (2018). Entropy, extremality, euclidean variations, and the equations of motion. Journal of High Energy Physics, 2018(1), 81