The application of principal component cluster analysis in environment classification for Chinese cities

Jingcheng Wang¹,²,³, Lunwu Zhang¹,³, Dingfei Zhang², Fangchao Zhao¹,³ and Xiaokui Yang¹

¹Southwest Technology and Engineering Research Institute, Chongqing 400039, China; ²College of Materials Science and Engineering, Chongqing University, 400044, China; ³CSGC Key Laboratory of Ammunition Storage Environmental Effects, China

Abstract. In order to investigate the dissimilarities of different cities in China, an approach combining principal component analysis and hierarchical clustering is proposed. Three rather than two principal components are reserved to conduct a more elaborate analysis. Based on corresponding component scores, dissimilarity between each city is measured during clustering. These cities are classified into seven types, and they are marked on the map of China. The result of this classification is consistent to our traditional cognition. Therefore, the principal component cluster analysis is suitable for analyzing numerous observations with variables on a large scale. This approach helps to enhance the environmental adaptability of equipments by recognizing the environment type of each city.

1. Introduction

With the coming data era, millions of data are generated every moment, and this poses tremendous challenge for analyzing. One of the most typical fields is meteorological environment [1]. There are over 60,000 meteorological stations all over China. Many environmental factors, like temperature, humidity, rainfall and atmospheric pressure, are widely monitored and recorded [2]. These valuable data resources may produce lots of benefits in economic, military, and social domains. Data analysis and processing is the key to making full use of these resources. With new data mining algorithm adopted to investigate the inherent rules among environmental factors, our cognition can be elevated to a higher level, and the implicit knowledge is further dug out.

Climate classification is originally from agricultural production. These classifications usually correspond to vegetation distribution in the sense that each climate type is dominated by one vegetation zone or eco-region [3]. Many approaches are developed to conduct this classification, among which Köppen climate classification and its modifications [3][4] are the most frequently adopted. Temperature and rainfall are the most commonly used variables for classification [5], because these two factors are the driven factors for vegetation distribution.

However, when it comes to equipments, more environmental factors such as humidity and solar radiation should be taken into consideration apart from temperature and rainfall. Solar radiation significantly accelerates the degradation of high polymer materials; and humid air easily causes the failure of electronic components. A comprehensive climate classification approach is urgently needed to act as guidance for the design and maintaining of different equipments.
This study aims to classify the environmental types of China according to equipment usage. Principal component analysis and hierarchy clustering are combined to carry out this investigation. Environmental data from 45 cities all over China is analyzed by the newly proposed method of principal component cluster analysis. 3 principal components are selected to represent 6 environmental variables, and based on their corresponding principal scores, dissimilarity between each city is measured.

2. Methods
Among all the problems encountered in data analysis, clustering is the most basic and common one. Data of observations often fall naturally into groups or clusters, where the characteristics of objects in the same cluster are similar and the characteristics of objects in different clusters are dissimilar. Zhang [6] employed fuzzy clustering to analyse the environment of atmospheric corrosion; Wu [7] conducted refinement research of atmospheric corrosion categories based on ordered sample clustering; Tang [8] studied the effects of chemical composition to the atmospheric corrosion by grey clustering analysis.

In data sets with many variables, groups of variables often move together. This is because more than one variable might be measuring the same driving principle which governs the behavior of the system. In meteorological systems, there are only a few such driving forces among lots of variables, so we can simplify the problem by replacing a group of variables with few new variables. Therefore, dimensionality reduction needs to be carried out to simplify analysis in meteorological data. As an efficient and widely applied algorithm[9-11], principal component analysis (PCA) gets rid of correlation factors from multi-variables, and selects several few independent factors, which can explain adequately for results. Adopting PCA, researchers are released from selecting variables and the computation is also simplified.

2.1. Principal component analysis
Principal component analysis (PCA) is a quantitatively rigorous method for achieving variable simplification. This method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. All the principal components form an orthogonal basis for the space of the data.

The first principal component is a single axis in space. When you project each observation on that axis, resulting values form a new variable. And the variance of this variable is the maximum among all possible choices of the first axis. The second principal component is another axis in space, perpendicular to the first one. Projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis. The full set of principal components is as large as the original set of variables. Normally, PCA follows these steps [9]:

First, an n-by-p data matrix \([X_{ij}]_{n \times p}\) is constructed. Rows of X correspond to observations and columns correspond to variables. Different variables are in different scales, so they need to be standardized as \(X'_j = (X_j - E(X_j))/\text{var}(X_j)\). Thus, variables are comparable, and \(X'_j\) forms a new matrix \([X'_{ij}]_{n \times p}\).

Second, the correlated coefficient matrix is calculated as following:
\[
R_y = \frac{\text{cov}(X_i, X_j)}{\text{var}(X'_j) \cdot \text{var}(X'_j)} = \text{cov}(X'_i, X'_j) \tag{1}
\]

Where, \(\text{cov}(X'_i, X'_j)\) is the covariance of \(X'_i\) and \(X'_j\).

Third, according to the eigen-equation \(|\lambda I - R| = 0\), work out the eigenvalues \(\lambda_i\) and their corresponding eigenvectors \(B_i\). Eigenvalues in descending order stand for component variance, and each eigenvector contains coefficients for one principal component.
Fourth, principal component variance contribution is calculated as:

\[ C_i = \frac{\lambda_i}{\sum_{m=1}^{n} \lambda_m}, \quad CC_k = \frac{\sum_{m=1}^{k} \lambda_m}{\sum_{m=1}^{n} \lambda_m} \quad (2) \]

Where, \( C_i \) stands for the percentage of the total variance explained by each principal component, and \( CC_k \) is the accumulated percentage of the first \( k \) components. Usually, the accumulated percentage is selected to be higher than 85% to make sure these components contain most of the variable information.

At last, principal component scores are determined by

\[ Z_{ij} = X_{ij}' \times B_i \quad (3) \]

2.2. Hierarchical clustering

Hierarchical clustering groups data over a variety of scales by creating a cluster tree. The tree is not a single set of clusters as in K-Means, but rather a multi-level hierarchy, where clusters at one level are joined as clusters at the next higher level. This allows researchers to decide the level or scale of clustering that is the most appropriate in application. To perform hierarchical cluster analysis on a data set, follow this procedure:

First, calculate the distance between observations in the data set to measure their similarity. The distance is often determined by Euclidean distance. There are many other ways to calculate this distance information, such as Chebychev distance, Cosine distance, Jaccard distance and Correlation distance.

Second, group the objects into a binary, hierarchical cluster tree. According to the similarity, pairs of observations in close proximity are paired into binary clusters. The newly formed clusters are further grouped into larger clusters until a hierarchical tree is formed.

Third, determine where to cut the hierarchical tree into clusters and assign all the observations below each cut into a cluster. These clusters can be created by specifying the number of clusters or inconsistency coefficient threshold.

3. Results and discussion

A data set including 6 environmental variables from 45 cities across China are collected to carry out our analysis. These latest ten-year (2010-2019) data are downloaded from Chinese meteorological data website. These variables are latitude \( X_1 \), annual average atmospheric pressure \( X_2 \), annual average temperature \( X_3 \), annual average relative humidity \( X_4 \), annual rainfall \( X_5 \), and annual sunshine time \( X_6 \). These environmental variables have significant influence on the performance of equipments. Environmental data of several cities are listed in Table 1. All the variables are not absolutely independent, but have intricate inherent relations. For example, temperature is apparently influenced by latitude, and humidity is also in relation to rainfall.

| Cities     | Latitude (°) | Pressure (kPa) | Temperature (℃) | Humidity (%) | Rainfall (mm) | Sunshine time (h) |
|------------|--------------|----------------|-----------------|--------------|---------------|-------------------|
| Beijing    | 39.8         | 101.3          | 13.4            | 51.8         | 578.8         | 2422.1            |
| Shanghai   | 31.4         | 101.6          | 17.5            | 68.6         | 1307.5        | 1465.3            |
| Chongqing  | 29.6         | 98.3           | 19.1            | 75.3         | 1231.4        | 1011.4            |
| Harbin     | 45.8         | 99.9           | 4.9             | 66.1         | 532.8         | 2083.6            |
| Zhengzhou  | 34.7         | 100.4          | 15.9            | 56.9         | 571.9         | 1832.8            |
| Guangzhou  | 23.2         | 100.5          | 21.9            | 78.2         | 2095.5        | 1611.1            |
Standardize these raw data in Table 1, then work out the correlated coefficient matrix $[R]$ (Table 2) using Eq.(1).

**Table 2. Correlation coefficient matrix**

|       | Latitude | Pressure | Temperature | Humidity | Rainfall | Sunshine time |
|-------|----------|----------|-------------|----------|----------|---------------|
| Latitude | 1.000    | -0.050   | -0.828      | -0.665   | -0.805   | 0.563         |
| Pressure | -0.050   | 1.000    | 0.512       | 0.452    | 0.432    | -0.489        |
| Temperature | -0.828 | 0.512    | 1.000       | 0.608    | 0.766    | -0.614        |
| Humidity | -0.665   | 0.452    | 0.608       | 1.000    | 0.859    | -0.805        |
| Rainfall | -0.805   | 0.432    | 0.766       | 0.859    | 1.000    | -0.728        |
| Sunshine time | 0.563 | -0.489   | -0.614      | -0.805   | -0.728   | 1.000         |

Based on the matrix in Table 2, eigenvalues and their corresponding eigenvectors are determined. In Table 3, it shows that when the first two principal components are counted, the accumulate contribution of variance reaches 85.41%; when the first three components are counted, this value is as high as 94.19%. Comparing with two principal components, 10.3% more information is reserved by three principal components.

**Table 3. Eigenvalues and corresponding eigenvector.**

| Principal component | Eigenvalue $\lambda$ | Contribution C (%) | Accumulate contribution CC (%) | Eigenvector B |
|---------------------|----------------------|--------------------|--------------------------------|---------------|
| 1                   | 4.143                | 69.05              | -0.403 0.268 0.429 0.440 0.462 -0.417 |
| 2                   | 0.981                | 16.36              | 0.547 0.811 -0.071 0.034 -0.105 -0.160 |
| 3                   | 0.527                | 8.79               | -0.217 0.307 0.630 -0.480 -0.066 0.475 |
| 4                   | 0.243                | 4.06               | 0.044 0.138 -0.251 0.396 0.475 0.730 |
| 5                   | 0.094                | 1.56               | 0.137 -0.028 -0.186 -0.621 0.721 -0.199 |
| 6                   | 0.012                | 0.19               | 0.686 -0.395 0.562 0.178 0.160 0.026 |

According to the first three eigenvectors, a 3-dimensional principal component loading is plotted in Figure 1. The constituents of principal components are visually presented.

The plot of two-dimensional component loading for 1-2 components, 1-3 components and 2-3 components are shown in Figure 2, 3, 4 respectively. It can be seen that temperature, humidity, rainfall and pressure are in positive correlation with the first principal component, whereas latitude and sunshine time are in negative correlation. In second principal component, while latitude, pressure and humidity are positively correlated, the remaining factors are negatively correlated.

Principal component scores are calculated by eigenvectors and variable values. In our study, the first three components are selected to cover main environmental information. Thus, six variables are reduced into three variables to identify different sites, but 94.19% information is reserved. In Figure 5, these scores are served as coordinates, so every city is assigned to a specific location in data space. In this three dimensional space, similarities of different cities are roughly shown. It can be seen that from present angle of view, Haikou-Wanning, Chongqing-Jiangjin and Changsha-Hangzhou are very close in data space, which means that they are similar in climate. However, it is not very convenient to recognize the similarity information directly from Figure 5. It often happens that from specific view angle, two cities are very close, but when the coordinate system is rotated, they become far from each
other. Therefore, in order to verify distances of every two cities, you need to spin the view over and over again. Another deficiency is that the similarities of different cities can’t be quantified by this way.

![Figure 1. The plot of three principal component loading.](image)

![Figure 2. The plot of 1-2 principal component loading.](image)

![Figure 3. The plot of 1-3 principal component loading.](image)

![Figure 4. The plot of 2-3 principal component loading.](image)

In order to quantify the similarity/dissimilarity directly, clustering analysis is employed after PCA. Use the principal component scores of different cities to perform hierarchy clustering. Hence, Euclidean distances are calculated to measure their similarity, and a hierarchy cluster tree is constructed in Figure 6. From the hierarchy tree, it can be determined that seven is the optimal cluster number to partition Chinese cities into different classifications. These seven clusters are cut from the hierarchy tree by a dash line with a value of 2.5. Below the dash line, dissimilarity of the cluster is lower than 2.5, and each cluster represents an environment type in China. According to the proposed PCCA approach, these 45 cities are classified into seven types, which are listed in Table 4. Chongqing-Jiangjin, Hanzhou-Changsha, Guilin-Nanning, Haikou-Wanning, Tianjin-Laoting are like twin cities from the standpoint of climate, since their values of dissimilarity are lower than 0.2. This indicates that they have almost the same climatic environment. The lowest value is 0.117 for Chongqing and Jiangjin. In fact, Jiangjin is a district in the municipality of Chongqing. Among these cities the highest dissimilarity value is 6.691, which is for Wanning and Golmud.
Figure 5. Principal component scores of different sites in data space.

Figure 6. Hierarchy clustering tree of Chinese cities.

The distribution of cities is plotted in Figure 7, cities of the same type are in the same color. South-east cities represented by Haikou are in yellow, which are hot and rain a lot. North-east cities are coloured in orange, these cold and high-latitude cities are represented by Mohe. North-west cities (green) represented by Dunhuang are featured by dry climate and little rainfall. Cities in North China Plain (blue) represented by Beijing are typical monsoon climate of medium latitude. Cities in Yangtze river basin (cyan) represented by Wuhan are typical subtropical monsoon climate, whose main features are warm and wet. Cities in Qinghai-Tibet Plateau (red) represented by Lhasa own the feature of low atmospheric pressure (high altitude) and long sunshine time. Cities in Yunnan-Guizhou Plateau (purple) are in low latitudes but high altitude, featured by warm climate and medium rainfall. This classification is consistent to our common sense.
Table 4. Environment classification of Chinese cities.

| Type 1       | Chongqing Jiangjin Hangzhou Changsha Wuhan Shanghai Hefei Nanjing Xian Chengdu Guiyang |
|--------------|----------------------------------------------------------------------------------------|
| Type 2       | Xichang Kunming                                                                      |
| Type 3       | Fuzhou Nanchang Xiamen Guangzhou Guilin Nanning Haikou Wanning                        |
| Type 4       | Beijing Tianjin Laoting Jinan Zhengzhou Taiyuan                                        |
| Type 5       | Erenhot Kumul Jiuquan Dunhuang Yinchuan Kashagar Turpan                                |
| Type 6       | Hailar Mohe Harbin Shenyang Changchun Altay                                          |
| Type 7       | Ganzhi Qamdo Lhasa Xining Golmud                                                     |

Figure 7. Clusters of Chinese cities marked in different colors.

4. Conclusions
Principal component cluster analysis approach is proposed to investigate climatic environment classification of Chinese cities. Based on principal component scores calculated by PCA, these cities are classified into seven types by hierarchical clustering. The result is basically consistent to our traditional cognition. Comparing with two dimensional PCA, three dimensional PCA combined with hierarchy clustering is more elaborate and efficient, since dissimilarities of different cities are quantified. Furthermore, with 10.3% more information, this method is especially suitable for analyzing numerous observations with variables on a large scale. In our following work, this approach will be further employed to investigate the worldwide environment classification. It helps to enhance the environmental adaptability of equipments by recognizing the environment type of each city.

Acknowledgement
This work was supported by the Chongqing Postdoctoral (231676) fund.

References
[1] Shen WH 2014 Analysis on big data application of meteorological data ICHINA 235 19
[2] Ma CL, Ma RP, Bai YH 2017 Characteristics of atmospheric environment in China’s coastal areas and atmospheric corrosion in typical coastal regions Equipment Environmental Engineering 14(8) 65

[3] Michal B, Eva H 2014 Climate classification revisited: from Köppen to Trewartha Climate Research 59(1) 1

[4] Peel MC, Finlayson BL, McMahon TA 2007 Updated world map of the Köppen-Geiger climate classification Hydrology & Earth System Sciences 11(3) 259

[5] Gerd S, Quirijn DJVL 2010 Computer assisted Köppen climate classification: a case study for Brazil International Journal of Climatology 27(2) 257

[6] Zhang LW, Wang XH, Feng XH 2003 The environment of atmospheric corrosion and environmental factors fuzzy clustering analyze. J. Chin. Soc. Corros. Prot. 23(2) 112

[7] Wu C, Fu DM, Li XG 2015 Refinement Research of atmospheric corrosion categories based on data analysis Equipment Environmental Engineering 12(4) 16

[8] Tang QH 2003 The effects of chemical composition to the atmospheric corrosion of steel-gained from grey clustering analysis J. Chin. Soc. Corros. Prot. 23(4) 239

[9] Jia ZJ, Zhao QL, Dong CF 2009 Evaluation of weathering conditions of 13 cities in China by principal component analysis J. Chin. Soc. Corros. Prot. 29(5) 388

[10] Fang HW, Sun SQ, Zhu YL 2009 Principal component analysis apply in water quality assessment Environmental Science and Management 34(12) 152

[11] Chen JH, Xie MY, Fu BQ 2006 Principal component analysis and cluster analysis of inorganic elements in panax quinquefolium Spectroscopy and Spectral Analysis 26(7) 1326