Study on the Application of PCA Data Analysis Method in Frontal Rainstorm Forecast

Pin Wang¹, Zongna Xiao², *

¹School of Economics, Guangzhou College of Commerce, Guangzhou 511363 China
²Department of Arts and Sciences, Guangxi Open University, Nanning 530022, China

*Corresponding author e-mail: 20180126@gcc.edu.cn

Abstract. According to the principle of weather dynamics, there are three principal components affecting rainstorm, namely ascending motion, unstable condition and water vapor condition. According to these three types, the forecast factor group of rainstorm is formed. The original multi-index is transformed into a few independent comprehensive indexes by principal component analysis (PCA). After standardization of various factor indexes, the comprehensive index of rainstorm forecast is obtained based on PCA. After all kinds of factors are standardized, the comprehensive index of rainstorm forecast is obtained based on PCA. The test results show that there is a good correspondence between the rainstorm and the forecast comprehensive index, and the trend basically corresponds to the peak and valley values. The forecast of the comprehensive index rainstorm has certain indication significance. Through the comparison between each classification index and independent samples, the threshold system of rainstorm forecast is established combined with comprehensive index and circulation classification, which has reference value for the presence or absence of rainstorm process forecast and is a beneficial discussion on the quantification of weather conceptual model forecasting method.

Keywords: Principal Component Analysis (PCA); Frontal rainstorm; Forecasting method.

1. Introduction

Principal component analysis (PCA) is to try to recombine many original indexes with certain correlation into a new set of linearly independent comprehensive indexes to replace the original indexes. PCA is a multivariate statistic method to examine that correlation between multiple variables, and study how to reveal the internal structure of multiple variables through a few principal component. That is, a few principal components are derived from the original variables, so that they retain as much information as possible from the original variables and are linearly independent of each other. Usually, the mathematical treatment is to use the original P indexes as a linear combination as a new comprehensive index.

Yufeng Li (2018) [1] took the air quality of 6 monitoring points as the research object, and analyzed the main influencing factors of the ambient air quality of each monitoring point based on PCA and fuzzy mathematics analysis. The results show that: 1) The air quality is judged to be Grade I by fuzzy
comprehensive evaluation; 2) The first principal component is air suspended particles and the second principal component is or; 3) The comprehensive score of ambient air quality of each monitoring point in different months is higher in January, February and December, and the pollution degree is more serious.

Taking some index data of environmental pollution and economic growth from 2003 to 2016 in Beijing, China as an example, Moyu Wang (2018) [2] analyzed economic factors based on PCA, and studies the relationship between economic development and environmental changes. Generally speaking, Beijing's economic growth and environmental pollution are generally "N-shaped", which indicates that when the economy is at a higher level of development, it will in turn inhibit the continuous deterioration of the environment.

The macro-environmental parameters involved in lightning occurrence are many and complicated. Without losing a lot of information, Xiaoyan Liu (2019) [3] compared the characteristics of two lightning weather processes based on the PCA. Using lightning, radar and weather background data, the lightning activity samples from Xining, China, from June to August 2011 to 2013 were selected, and 26 environmental parameters from the same sounding data were extracted for PCA to deeply analyze the evolution and environmental conditions of the two lightning weather processes on July 29th, 2012 (Process I) and August 26th, 2013 (Process II). The results show that the parameters of atmospheric dynamics and temperature and humidity are most closely related to lightning activity in Xining, China. The cumulative probability of lightning intensity shows that the dynamic type is stronger than the temperature and humidity type, the average negative ground flash intensity is higher than -1.91 kA, and the average positive ground flash intensity is higher than 1.92 kA. Process I is characterized by westerly trough lightning weather, which is greatly affected by atmospheric temperature and humidity parameters. However, Process II belongs to subtropical high edge lightning weather, which is obviously affected by atmospheric dynamic parameters. The lightning activity of the two processes is mainly negative ground flash, which has obvious diurnal variation law and is distributed in a unimodal pattern. Process I is dominant in lightning frequency, but obviously weaker than Process II in intensity.

At present, the forecast of rainstorm mainly depends on conceptual model and statistical forecasting method. Statistical forecasting methods are widely used in regression, similarity and other forecasting methods. The essence of these statistical forecasting methods is mostly linear sequential processing technology. However, due to the complex influence mechanism of rainstorm weather process, the stability of deterministic forecasting is poor. The conceptual model prediction method has a clear analysis idea of synoptic principles, but it is more qualitative circulation classification and index groups, which makes it difficult for forecasters to popularize and apply it due to its large randomness in actual use. In practical application, the basic physical quantities that play an important role in the occurrence and development of rainstorm can be calculated quantitatively or qualitatively, and whether rainstorm weather would occur can be predicted from the matching mode of these basic quantities. This idea has been applied to the international severe convective weather forecast. This paper attempts to use PCA to transform the original multi-index into a few independent new factor groups. Based on PCA, the comprehensive index of rainstorm forecast is obtained after all kinds of factors are standardized, and the threshold system of rainstorm forecast is established through comparison between each classification index and independent samples, combination of comprehensive index and circulation classification, which has good reference value for the presence or absence of rainstorm process forecast and is a beneficial discussion on the quantification of weather conceptual model forecasting method.

2. Data Principal Component Analysis

2.1. Analysis of basic theory of rainstorm weather process

On the basis of understanding the physical process of rainstorm occurrence, the analysis of the basic factors causing the formation of rainstorm weather refers to the method of forecasting the possibility of precipitation from the viewpoint of basic components, which is suitable for forecasting rainstorm and severe convective weather. "Composition" is defined as a basic physical quantity or process that directly
affects the development and intensity of precipitation events. The correct combination of basic components can produce predicted weather phenomena, which provides fixed-point, regular and quantitative forecasts. For rainstorm, several basic physical components are upward movement, unstable conditions and water vapor conditions. In practical application, the corresponding diagnostic quantity can be selected to represent it. There are many diagnostic factors. How to select appropriate diagnostic factors to characterize the occurrence and development of rainstorm is the key to establish a mathematical model.

2.2. Principal components in rainstorm forecast
A Based on the classification of frontal rainstorm in Guangxi [4], the ascending motion, unstable conditions and water vapor conditions of different rainstorm types are further analyzed. The research results show that the following diagnostic parameters have good corresponding relations with the occurrence and development of frontal rainstorm in Guangxi, which can be used as the basic principal components of rainstorm forecast.

2.2.1. Water vapor condition factor. ① Relative humidity of the bottom layer; ② Water vapor flux of the lower water vapor flow person; ③ Relative humidity of water vapor in the middle and higher layers; ④ Water vapor flux of water vapor flow in middle and higher layers.

2.2.2. Uplift conditions. ① Lower vorticity ② Vertical velocity ③ Higher divergence ④ North-south wind index in middle and lower layers

2.2.3. Stability. ① K index ② Convective Available Potential Energy (CAPE)

3. Processing of Forecast Factors
This paper selects some cases of frontal rainstorm process in Guangxi from 2001 to 2011. Frontal rainstorm process is defined as more than 10 cities and counties rainstorm in Guangxi within 24 hours. The total number of samples is 105, of which 85 samples are used as statistical calculation and the other 20 samples are used as independent samples for prediction analysis. According to the synoptic principles and many years of practical experience in forecasting, 14 forecasting factors are selected through correlation analysis methods, including ascending motion, unstable conditions and water vapor conditions. These forecast factors have good correlation with the forecast quantity, i.e. rainstorm weather, and the absolute values of correlation coefficients are between 0.41 and 0.65, and all reach the correlation significance level of 0.05, which indicates that the selected forecast factors have good correlation with the forecast quantity.

3.1. Factor standardization treatment
After the establishment of the factor system, because the dimensions of each factor are not unified and there is no comparability between factors, all data must be dimensionless to make each data standardized and comparable. The calculation method is as follows:

1) Positive effect index
The standard values for the i-th sample of the factor are:

\[ P_y = \frac{P_y' - \min(P_y')}{\max(P_y') - \min(P_y')} \quad (1) \]

2) Negative effect index

\[ P_y = \frac{\max(P_y) - P_y}{\max(P_y) - \min(P_y)} \quad (2) \]
In (1) and (2), $P_{ij}$ is the normalized factor value, $P_{ij}'$ is the original factor value, max is the maximum value, and min is the minimum value.

### 3.2. Recombination of factor groups by PCA

PCA is a multivariate statistical analysis method that uses the idea of dimension reduction to transform the original multiple indexes into a few independent comprehensive indexes [5], which is widely used in the prediction and analysis of atmospheric science. The basic idea is that if a forecast object has $n$ forecast factors, $X = (x_1, x_2, \cdots, x_n)^T$, $n$ new comprehensive factor variables can be constructed by PCA, $Z = (z_1, z_2, \cdots, z_n)^T$. Each new comprehensive factor variable is a linear combination of cause subvariables, and the new factor variables are orthogonal to each other, that is, the correlation coefficient between each factor is zero, and the eigenvalue of each new comprehensive factor variable clearly indicates its contribution to the cause subgroup.

**Table 1.** Eigenvalues of principal components and correlation coefficients between principal components and forecast variables

| No. | Eigenvalue | 1   | 2   | 3   | 4   | 5   | 6   | 7   |
|-----|------------|-----|-----|-----|-----|-----|-----|-----|
|     |            | 6.57| 2.95| 1.44| 1.01| 0.84| 0.75| 0.50|
| Correlation coefficient | 0.56 | 0.49| -0.28| 0.15| 0.01| 0.13| 0.04|
| No. | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Eigenvalue | 0.31 | 0.11| 0.09| 0.03| 0.00| 0.00| 0.00|
| Correlation coefficient | 0.17 | 0.02| -0.04| 0.14| -0.20| -0.12| 0.08|

The above prediction factors are analyzed by PCA. Table 1 is the eigenvalues of each principal component and the correlation coefficients between each principal component and the forecast quantity calculated from the primary forecast factor group as the principal component. The new index group is determined by taking the relatively large eigenvalue of the principal component and considering the good correlation with the forecast quantity as the selection standard. As can be seen from Table 1, relatively speaking, the absolute value of correlation coefficient between the first, second, third and fourth principal components and the forecast quantity is the largest, and the corresponding eigenvalues are also relatively large. Therefore, these four principal components are used to construct independent new factor groups, which are respectively recorded as $P_1$, $P_2$, $P_3$ and $P_4$.

### 4. Calculation of Comprehensive Index of Rainstorm Forecast

#### 4.1. Determination of weight and construction of comprehensive index of rainstorm forecast

The eigenvalue of the factor variable clearly indicates its contribution to the cause subgroup, so we measure the role of the factor in the comprehensive evaluation by the eigenvalue. If the eigenvalue of the factor is larger, the greater its role in the comprehensive evaluation is, and vice versa. Therefore, we calculate the factor weight value and the comprehensive index of rainstorm forecast based on the eigenvalue. The equation is

$$H_i = 100 \times \sum_{j=1}^{q} W_i P_{ij}$$

$$h_j = 100 \times W_j P_{ij}$$

$$W_i = \frac{\sum_{j=1}^{q} O_j}{Q_j}$$
In the equation, $H_i$ is the comprehensive index of rainstorm forecast, $P_{ij}$ is the standardized factor quantity value, and $W_i$ is the weight of factor $P_i$, $Q_i$ is the eigenvalue.

### Table 2. Weights of New Factor Groups

|     | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
|-----|-------|-------|-------|-------|
| Weight | 0.55  | 0.24  | 0.12  | 0.08  |

According to Equation (5), 85 samples are used to calculate the weight of each new index group (Table 2). From Table 2 and the standardized values of each sample, the rainstorm forecast comprehensive index $H_i$ of 20 samples that did not participate in statistical analysis is calculated according to Equation (3), and compared and analyzed with the number of rainstorm process stations (see Figure 1).

5. Analysis of Calculation Results

5.1. Comparative analysis of forecast comprehensive index and rainstorm station number

As can be seen from Figure 1, the number of rainstorm stations has a good corresponding relationship with the forecast comprehensive index, and the trend basically corresponds to the peak and valley values. It can be seen that the comprehensive index has a good indicating significance for the forecast of rainstorm process. There are more than regional rainstorm processes in Guangxi, and most of the comprehensive index values are greater than 30. The comprehensive index value of 30 can be taken as the starting threshold for forecasting rainstorm processes.

5.2. Correspondence analysis between index of each new factor group and number of rainstorm stations

According to Equation (4), the exponent of each new factor group is calculated as $h_i = W_i P_{ij} (i=1,2,3,4)$. Figure 2 is a correspondence analysis chart between the index of the first two factor groups and the number of rainstorm stations. From Figure 2, it can be seen that the index of the first factor group has a good positive correlation with the number of rainstorm stations, and the trend basically corresponds to the peak and valley values, while the index of the second factor group has a poor correlation with the number of rainstorm stations. Because the new factor groups are independent of each other and do not
cross each other, we can find a finer corresponding relationship between rainstorm weather and the indexes of each new factor group through circulation typing, and establish the ingredient combination of Guangxi frontal rainstorm and the indexes of each new factor group.

In addition, in the trial report of the first flood season in 2012, the comprehensive rainstorm forecast index value of 0.30 is taken as the starting report threshold for forecasting the rainstorm process in Guangxi front with more empty reports and less missing reports, which indicates that the index can be used as the starting report threshold for the presence or absence of rainstorm process. If the accuracy rate is to be improved, it is also necessary to eliminate empty reports in combination with circulation classification.

![Figure 2](image)

**Figure 2.** Evolution curves of new factor indexes I1 (solid circular line, and right ordinate), I2 (hollow square line and right ordinate) and number of rainstorm stations (hollow circular line and left ordinate)

### 6. Conclusions

1. According to the principle of synoptic dynamics, three kinds of factors affecting the rainstorm process in Guangxi frontal are obtained, namely, ascending motion, unstable condition and water vapor condition. Based on years of experience, gale forecasting factor groups are formed according to these three types. The original multi-index is transformed into a few independent new factor combinations by PCA.

2. After all kinds of factor indexes are standardized, the comprehensive index of rainstorm forecast and the index of each new factor group are obtained based on PCA.

3. The test results show that the number of rainstorm stations has a good correspondence with the comprehensive index of rainstorm forecast, and the trend basically corresponds to the peak and valley values, which shows that the comprehensive index has a good indication significance for rainstorm forecast. The comprehensive index value of 0.30 can be taken as the starting threshold for forecasting strong winds in Beibu Gulf. It has a good reference value for forecasting the presence or absence of rainstorm and is a useful discussion for quantifying the forecasting method of weather conceptual model.

4. We can find the more precise correspondence between rainstorm weather and the indexes of each new factor group through circulation typing, and establish the principal component combination of Guangxi frontal rainstorm and the indexes of each new factor group.

5. The comprehensive index of rainstorm forecast can be used as the starting threshold for the presence or absence of rainstorm process, but there are many empty reports. If the accuracy rate is to be improved, it is necessary to eliminate empty data in combination with circulation classification.

**Acknowledgments**

This research work was supported by Guangdong Higher Education Teaching Revolution Project:
Experiencing Teaching and Mode Design of Cultivation of Internet Finance Talents under the Background of Financial Science Technology, project number: 2018SJJXGG01; Guangzhou College of Commerce Higher Education Teaching Revolution Project: Teaching Reform and Practice of Securities Investment under the Cultivation Method of Applied Talents, project number: 2018XJJXGG01. Foundation projects: Guangdong Undergraduate Higher Education Teaching Quality Engineering project: “Guangzhou College of Commerce -Applied Training Base of Mobile Commerce and E-commerce Logistics, project number: SJJD201501; Guangdong College Characteristic Innovative Project (humanities and social sciences category): “Virtual Reality Experimental Teaching Center of Guangzhou College of Commerce”, project number: 2015WTSCX114.

References

[1] Yu-feng Li, Fuzhou City Air Quality Assessment Based on the Principal Component Analysis and Fuzzy Comprehensive Evaluation, J. Journal of Wuyi University. Vol.37, No.9 (2018) 39-44.

[2] Moyu Wang, Research on the Relationship between Beijing's Economic Development and Environmental Pollution Based on Principal Component Analysis, J. Electronic Technology & Software Engineering. 23 (2018) 182-183.

[3] Xiao-yan Liu, Comparation on Characteristics of Two Lightning Processes Based on the Principal Component Analysis, J. Meteorological and Environmental Sciences. Vol.42, No.4 (2019) 119-126.

[4] Jing Li, Conceptual Model of the Frontal Storms in Guangxi, J. Journal of Meteorological Research and Application. Vol.31, No.4 (2010) 8-11.

[5] Ji-fu Nong, The Rainfall Forecast Model of PCA-RBF Neural Networks Based On MATLAB, J. Journal of Tropical Meteorology. Vol.24, No.6 (2008) 713—717.