*Advanced Engineering Forum* *Online: 2012-09-26*
*ISSN: 2234-991X, Vols. 6-7, pp 676-681*
*doi:10.4028/www.scientific.net/AEF.6-7.676*


# The Application of the Grey Correlation Method in the Principal Component Analysis

Baoping Chen

Department of Computer Information and Management NeiMongol University of Finance and Economics, Hohhot, China

e-mail:cbpmoonsong@126.com

**Keywords:** Gray correlative analysis, principal component analysis, compositive correlative degree

**Abstract** To solve the shortcoming of the traditional principal component analysis in term of the processing of nonlinear data, this paper puts forward one principal component analysis method based on the grey correlation method. In face of the multiple indexes comprehensive evaluation problems, this paper establishes the evaluation model by aid of the grey correlation method and by calculating the comprehensive correlation degree of each evaluation index to find out the major factors affecting the problems. This paper aims at apply this method to evaluate the development of the telecommunication industry in various regions all over the country and to explore the major reason for the differences. The results show that this method can comprehensively considerate the various factors of the evaluation problem, which not only avoids the subjectivity of the single factor but also makes the analysis process more reasonable and objective. Also the analysis results can accurately reflect the differences between various factors. Gray correlation analysis method has a low requirement on the linearity and regularity of the original data and there is no case where the quantitative result is inconsistent with the qualitative analysis result.

## Introduction

Gray correlation analysis method is one method to quantitatively describe as well as compare the trend of the development and change of a system[1].Through the determination of the similarity degree of the geometrical shapes between the reference data column and several comparative data columns, the closeness of the correlation is estimated and the correlation degree of curves also is reflected. In the development process of one dynamic system, the major influential factors can be analyzed through the sequencing of the correlation degrees. Among them, a low correlation degree means there is no or less influence from this factor while a high correlation degree means this factor is the major one influencing the development of the system.

The principal component analysis is one statistical analysis method which converts multi indexes into several comprehensive indexes as well as a method which converts multiple variables into several principal components. The principal components retain the most of the information about the original variables. Through the principal component analysis, some principal components can be found from the complicated relationships of things. In this way, a large number of statistical data can be effectively used to carry out the quantitative analysis and further the internal correlation between variables can be revealed, the underlying inspiration of the characters of things and the law of development can be obtained[2]. All these lead the research into in-depth study. Traditional principal component analysis is a linear dimension reduction technique. However, in the practical application, the indexes of the original data usually present nonlinear relationships and the dimension reduction effect of the principal component analysis is not ideal. Sometimes the results even show a large deviation between the evaluation and the fact. Therefore, this paper adopts grey correlation method to carry out the principal component analysis. Grey correlation method is a kind of multi-factor statistical analysis method and owns obvious advantages in case of inaccurate information and not completely certain small-sample system[3]. There are quite a lot of researches on the principal component analysis in China, but most of them adopt the traditional method, which is different from the method discussed in this paper. In face of the multiple indexes compre-

hensive evaluation problems, this paper adopts gray correlation analysis method[4]. Through the calculation of the absolute correlation degree, relative correlation degree and comprehensive correlation degree of the index, the major factors affecting the problem can be obtained. The data sample in the evaluation can be very small and the index of the original data can be nonlinear, so this method can have a wider applicability in solving most problems.

With the rapid development and change in the information age, the telecommunication service has been one of the growth points of China's national economy and the most important pillar industries. However, the developments of telecommunication industry in different provinces and cities in China are uneven and the differences between regions are great. Specific to the development of telecommunication industry in each region, this paper adopts grey correlation method to explore the major reasons for the differences and find out the solutions to the problems. In this way, the common development of each region can be realized and the fall of the overall level due to certain less developed regions can also be avoided. Tested by practice, the grey correlation method is an effective method to carry out the principal component analysis on things.

**Procedures of grey relation analysis**

Grey correlation method has a low requirement on the sample size and its regularity. It can be applied to the evaluation study with few statistical data, large data grey, great data fluctuation or non-typical distribution regularity. Grey correlation method based on the gray system theory is a multi-factor analysis technique [5] which uses grey correlation to describe the strength, degree and sequence of the correlation between the factors through the calculation of the grey correlation degree. The specific procedures are as follows.

Step 1: determine the reference sequence and comparative sequence. On the basis of the qualitative analysis, determine one dependent variable and multiple independent variables. For m indexes which have n evaluation objects, according to the historical statistics, the reference sequence X0 which reflects the corresponding condition of the things and the comparative sequence which describes the corresponding situation of m factors are given. Among them:

Reference sequence:

$$X_0 = （x_0(1),X_0(2), \cdots X_0(m)） \tag{1}$$

Comparative sequence:

$$X_i = (x_i(1),x_i(2),...,x_i(m)) \quad （i=1,2,…,n） \tag{2}$$

Step 2: calculate the absolute correlation degree. Set the length of Xi is same as that of Xj and Xi0 and Xj0 are their respective initial point zero images.

$$X_i^0 = (x_i(1)-x_i(1), x_i(2)-x_i(1),..., x_i(m)-x_i(1))) \text{ (i=0,1,2,…,n)} \tag{3}$$

Formula:

$$\varepsilon_{ij} = \frac{1+|s_i|+|s_j|}{1+|s_i|+|s_j|+|s_i-s_j|} \tag{4}$$

Calculate the grey absolute correlation degrees of Xi and Xj. Among them:

$$|s_i| = \left|\sum_{k=2}^{n-1} x_i^0(k) + \frac{1}{2}x_i^0(n)\right| \tag{5}$$

Step 3: calculate the relative correlation degree. Set the length of $X_i$ is same as that of $X_j$ and their initial values are not equal to zero. $X_i'$ and $X_j'$ are respectively the initial value images of $X_i$ and $X_j$. Take $\varepsilon_{ij}'$, the absolute correlation degree of $X_i'$ and $X_j'$ as the grey relative correlation degree of $X_i$ and $X_j$. Note as $r_{ij}$. Among them:

$$X_i' = X_i / x_i(1),\ X_j' = X_j / x_j(1) \tag{6}$$

$$r_{ij} = \frac{1+|s'_i|+|s'_j|}{1+|s'_i|+|s'_j|+|s'_i - s'_j|} \tag{7}$$

Step 4: solve the comprehensive correlation degree. Set the length of $X_i$ is same as that of $X_j$ and their initial values are not equal to zero. There is no positive correlation between $\varepsilon_{ij}$ (the absolute correlation degree of $X_i$ and $X_j$) and $r_{ij}$ (the relative correlation degree of $X_i$ and $X_j$). Comprehensive correlation degrees take both the absolute change and relative change of the data sequences into consideration and at the same time satisfy 4 Axiom of grey correlation degree [6]. Note $\rho_{ij}$ as the grey comprehensive correlation degrees of $X_i$ and $X_j$. Among them:

$$\rho_{ij} = \theta\varepsilon_{ij} + (1-\theta)r_{ij} \qquad \theta \in [0,1] \tag{8}$$

The value of θ represents the emphasis on the absolute correlation degree ε ij and relative correlation degree rij. Generally, the value of θ is 0.5. When θ is set, grey comprehensive correlation degree is unique. However, this kind of conditional uniqueness does not affect the analysis on the problem.

Step 5: utilize the calculated comprehensive correlation degree analysis $\rho_{ij}$ to analyze the correlation sequence.

**Case studies**

At present, telecommunication services have accomplished a series of conversions, from the manual work to automation, from simulation technique to digital technique, from small capacity to large one and from single business volume to multiple ones. It has been one of the growth points of China's national economy and the most important pillar industries. In 2010, in the increasingly fierce market competition, telecom companies around China took forceful measures to strive for the market share in succession and afforded the telecom users with more benefits. Telecommunication service keeps a rapid development and its comprehensive strength has reached a new height. However, the developments of telecommunication industry in different provinces and cities in China are uneven and the differences between regions are great. This paper adopts grey correlation method to explore the major reasons for the differences and find out the solutions to the problems. In this way, the common development of each region can be realized and the fall of the overall level due to certain less developed regions can also be avoided.

The provinces, municipalities and autonomous regions all over the country are divided into three parts. The eastern region includes Beijing, Hebei, Liaoning, Shanghai, Jiangsu, Zhejiang, Shandong, Guangdong, etc.; the central region includes Shanxi, Heilongjiang, Henan, Hubei, etc.; the western region includes Yunnan, Guangxi, Qinghai, Ningxia, Xinjiang, etc. According to the Chinese Statistic Almanac of 2010 which provides the development data of telecommunications industry of these 24 regions in 2010, this paper takes the gross telecommunication services as the reference sequence and 7 major indexes (including local fixed-line calls, call-time of long-distance fixed-line calls, call-time of mobile calls, call-time of IP calls, mobile short message service, internet-surfing population and year-end users of mobile calls) as the evaluation indexes. Table 1 shows the relevant data of the 24 cities in 2010.

From table 1, the reference sequence can be obtained:

$X_0$=(399.46, 1293.51, 696.69, 584.75, 1113.22, 587.4, 697.87, 1098.79, 2140.42, 1947.77, 841.62, 1145.33, 655.15, 1855.86, 1383.65, 983.32, 1008.94, 4175.38, 793.31, 214.76, 111.07, 131.1, 537.32, 1120.0）

Comparative sequence is:

$X_i$=（$x_i(1)$, $x_i(2)$, $x_i(3)$,…, $x_i(24)$） （i=1,2,…,24）

(1) Calculate the absolute correlation degree. Take the local fixed-line calls for example, through the initialization operation (settled as 1- time interval sequence of equal length), we can obtain:

$X_1$=(68.0,173.2,103.2,52.9,217.4,83.6,116.5,229.3,276.9,228.2,141.9,150.7,80.4,256.8,295.5,105.2, 145.6,565.7, 151.3,27.6,14.0,13.4,98.2,214.8)

Through the operation of initial point zero images on X0 sequence and X1 sequence, we can obtain following sequences:

$X_0$=(0.0000, 894.0500, 297.2300, 185.2900, 713.7600, 187.9400, 298.4100, 699.3300, 1740.9600, 1548.31, 442.16, 745.8700, 255.69, 1456.4, 984.19, 583.8600, 609.48, 3775.92, 393.85, -184.7, -288.39, -268.36, 137.86, 720.54)

Table 1 The relevant data of the 24 cities in 2010

| regions | the gross telecommunication services | local fixed-line calls | long-distance fixed-line calls | call-time of mobile calls | call-time of IP calls | mobile short message service | internet-surfing population | year-end users of mobile calls |
|---|---|---|---|---|---|---|---|---|
| Tian jin | 399.46 | 68.0 | 7.3 | 548.9 | 23.5 | 124.9 | 648 | 1089.8 |
| He bei | 1293.51 | 173.2 | 24.9 | 2091.2 | 5.1 | 352.5 | 2197 | 4353.6 |
| Shan xi | 696.69 | 103.2 | 15.2 | 1041.3 | 4.4 | 219.8 | 1250 | 2205.2 |
| Nei menggu | 584.75 | 52.9 | 8.5 | 1032.9 | 5.4 | 181.9 | 747 | 2034.0 |
| liao ning | 1113.22 | 217.4 | 24.8 | 1578.8 | 22.4 | 257.9 | 1916 | 3341.8 |
| Jie lin | 587.40 | 83.6 | 9.0 | 1002.6 | 3.9 | 177.5 | 882 | 1805.4 |
| Hei longjiang | 697.87 | 116.5 | 14.5 | 1201.2 | 12.0 | 177.6 | 1127 | 2072.0 |
| Shanghai | 1098.79 | 229.3 | 35.8 | 1011.0 | 129.3 | 369.0 | 1239 | 2361.6 |
| Jiang su | 2140.42 | 276.9 | 58.2 | 2735.7 | 24.4 | 680.8 | 3306 | 5923.1 |
| Zhe jiang | 1947.77 | 228.2 | 57.0 | 2660.2 | 16.6 | 621.6 | 2786 | 5047.4 |
| An hui | 841.62 | 141.9 | 13.5 | 1104.5 | 37.9 | 317.2 | 1392 | 2798.7 |
| Fu jian | 1145.33 | 150.7 | 25.7 | 1663.1 | 19.7 | 306.6 | 1848 | 3021.8 |
| Jiang xi | 655.15 | 80.4 | 12.6 | 1022.4 | 7.0 | 160.5 | 950 | 1811.3 |
| San dong | 1855.86 | 256.8 | 29.0 | 2771.8 | 46.4 | 447.8 | 3332 | 6190.4 |
| He nan | 1383.65 | 295.5 | 29.2 | 2177.6 | 7.4 | 317.2 | 2417 | 4402.0 |
| Hu bei | 983.32 | 105.2 | 32.1 | 1327.3 | 37.8 | 253.5 | 1902 | 3454.7 |
| Hu nan | 1008.94 | 145.6 | 22.3 | 1593.0 | 34.9 | 251.9 | 1747 | 3257.0 |
| Guang dong | 4175.38 | 565.7 | 149.9 | 5536.0 | 221.3 | 931.5 | 5324 | 9624.6 |
| Guang xi | 793.31 | 151.3 | 23.9 | 1167.7 | 7.1 | 188.2 | 1226 | 2214.5 |
| Hai nan | 214.76 | 27.6 | 5.4 | 375.5 | 4.3 | 49.5 | 303 | 594.3 |
| Qing hai | 111.07 | 14.0 | 4.3 | 159.6 | 2.1 | 27.5 | 188 | 397.8 |
| Ning xia | 131.10 | 13.4 | 2.8 | 220.0 | 1.2 | 44.1 | 175 | 437.3 |
| Xin jiang | 537.32 | 98.2 | 15.4 | 885.9 | 53.5 | 96.4 | 819 | 1359.9 |
| Bei jing | 1120.00 | 214.8 | 26.8 | 1163.6 | 72.0 | 369.6 | 1218 | 2129.8 |

$X_1$=( 0.0,105.2,35.2,-15.1,149.4,15.6,48.5,161.3,208.9,160.2,73.9, 82.7,12.4,188.8, 227.5,37.2,77.6,497.7,83.3, -40.4,-54.0,-54.6,30.2,146.8)

Calculate the values of $|s_0|$, $|s_1|$ and $|s_1 - s_0|$ . Among them:

$|s_0| = 15569.38$ ； $|s_1| = 2104.9$； $|s_1 - s_0| = 13464.48$

Thus, according formula (4), the absolute correlation degree of local fixed-line calls can be calculated and its value is 0.5676. Similarly, the absolute correlation degrees of all the factors can be calculated. Namely:

$\varepsilon_{01}$=0.5676， $\varepsilon_{02}$=0.5149， $\varepsilon_{03}$=0.8446，

$\varepsilon_{04}$=0.5068， $\varepsilon_{05}$=0.6222 ， $\varepsilon_{06}$=0.8370, $\varepsilon_{07}$=0.6720

(2) Calculate the relative correlation degree. Take the local fixed-line calls for example. After the initialization operation, calculate the initial value images of $X_0$ sequence and $X_1$ sequence. Namely:

$X_0$=(1.0,3.2381,1.7441,1.4639,2.7868,1.4705,1.7470,2.7507,5.3583,4.8760,2.1069,2.8672,1.64,4.6459,3.4638,2.4616,2.5258,10.4526,1.9860,0.5376,0.2781,0.3282,
1.3451,2.8038)

$X_1$=(1.0,3.411,2.0822,1.1644,3.3973,1.2329,1.9863,4.9041,7.9726,7.8082,1.8493,
3.5205,1.7260,3.9726,4.0, 4.3973,3.0548,20.5342,3.2740,0.7397,0.5890,0.3836,
2.1096,3.6712)

Calculate the initial point zero images of $X_0$ ′and $X_1$′. Namely:

$X_0$′=(0.0,2.2381,0.7441,0.4639,1.7868,0.4705,0.7470,1.7507,4.3583,3.8760,
1.1069,1.8672,0.6401,3.6459,2.4638,1.4616,1.5258,9.4526,0.9860,-0.4624,
-0.7219,-0.6718,0.3451,1.8038)

$X_1$′=(0.0,2.4110,1.0822,0.1644,2.3973,0.2329,0.9863,3.9041,6.9726,6.8082,
0.8493,2.5205,0.7260,2.9726,3.0,3.3973,2.0548,19.5342,2.2740,-0.2603,-0.4110,
-0.6164,1.1096,2.6712)

The values of $|s_0|$, $|s_1|$ and $|s_1 - s_0|$ can be obtained.

$$|s_0| = 38.9762 \; ; \; |s_1| = 3095 \quad .47 \; ; \; |s_1 - s_0| = 8.0215$$

Thus, according formula (7), the relative correlation degree of local fixed-line calls can be calculated and its value is 0.8984. Similarly, the relative correlation degrees of all the factors can be calculated. Namely:

$r_{01}$=0.8984, $r_{02}$=0.8087, $r_{03}$=0.9738, $r_{04}$=0.6205, $r_{05}$=0.8922，$r_{06}$=0.9579，$r_{07}$=0.9697

(3) Calculate the comprehensive correlation degree. Utilize the above absolute correlation degree and relative correlation degree and formula (8) and at the same time set $\theta$ =0.5, the comprehensive correlation degrees of all the factors can be calculated. Namely:

$\rho_{01}$=0.7330， $\rho_{02}$=0.6618， $\rho_{03}$=0.9092， $\rho_{04}$=0.5636， $\rho_{05}$=0.7572，

$\rho_{06}$=0.8974, $\rho_{07}$= 0.8209

(4) Result analysis.

The result is $\rho_{03} < \rho_{06} < \rho_{07} < \rho_{05} < \rho_{01} < \rho_{02} < \rho_{04}$,， Namely: $X_3 \prec X_6 \prec X_7 \prec X_5 \prec X_1 \prec X_2 \prec X_4$。

$X_4$ is the optimal factor, namely, the major factor affecting telecommunication services is call-time of mobile calls. Internet-surfing population, year-end users of mobile calls, mobile short message service and local fixed-line calls come after. The factors with least influences are call-time of long-distance fixed-line calls and call-time of IP calls. These results are consistent with the fact. With the rapid development of telecom technology and increasingly intense market competition in telecommunication industry, if the telecom operators want to improve the gross telecommunication services, importance should be attached to the users of mobile phone and Internet by affording the corresponding preference policies to attract more users and creating the greatest benefit for them. At the same time the service quality and efficiency should be strengthened.

## Conclusions

Grey comprehensive evaluation method is a comprehensive evaluation method which combines the qualitative analysis and quantitative analysis. This method can not only solve the problems of evaluation indexes well that the evaluation indexes are difficult to quantify and accurately statistic, but also exclude the effects of personal factors. All these make the evaluation results more accurate. Gray correlation analysis method adopts the correlation degree to quantitatively describe the strength of the influences between things[6]. The calculated value of the correlation degree falls on the interval [0, 1]. The larger the value is, the stronger the influence between things is. The geometric significance of the correlation degree is the difference degree of the geometrical shapes between curves which represent different things or factors. If the correlation degree of certain index is high, it means this index is one major factor affecting things. On the contrary, if the correlation degree of certain index is low, it means this index has a low influence. Applying grey correlation method into the principal component analysis to seek the major influential factors can take several factors into consideration comprehensively, which avoids the subjectivity of the single factor. In this way, the analysis process can be more reasonable and objective and the analysis results can accurately reflect the differences between various factors. The above case shows that gray correlation analysis method has a low requirement on the regularity of the original data and definite objectivity and scientificalness. Besides, it is simple for use, not time consuming and easy to understand.

**References**

[1] Dang Yaoguo Liu Sifeng&Wang Zhengxin . (2009) Model on gray prediction and decision model .Beijing: Science Press.

[2] Wang Huiwen，Li Yan&Guan Rong.( 2011) A Comparison Study of Two Methods for Principal Component Analysis of Interval Data. Journal of Beijing University of Aeronautics and Astronautics( Social Sciences Edition), 7 ,86-89

[3] FANG Fang，TANG Wu-xiang & CHENG Gui-zhi. (2011) Performance evaluation of beijing innovative enterprises based on principal component analysis . Journal of Beijing Information Science and Technology University,8,89-94

[4] ZHANG Jingyu，HU Xiaohua，LIN Xiao. (2011)Research on The Financial Revenue of Hainan ProvinceBased on the Principal Component Analysis. Journal of Hainan Normal University（Natural Science）, 9,260-264.

[5] Liu Sifeng&Dang YaoGuo . (2009)Grey System on dealing with the theory and pratical applications .Beijing: Social Sciences Edition.

[6] Sun Lei . (2011)Comparison between Performance of Principal Component Analysis and Fuzzy Analysis in Water Quality Evaluation. ENVIRONMENTAL SCIENCE AND MANAGEMENT,8,178-181.