International Core Journal of Engineering Volume 8 Issue 8, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0080

Comprehensive Evaluation Model based on Principal Component Analysis and Cluster Analysis

Maojie Liao, Zhengting He, Qi Liu

School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China

These authors contributed equally to this work.

Abstract

Firstly, the principal component comprehensive evaluation model and impact assessment model are established to verify the important impact of Saihan dam restoration on its ecological environment and the huge ecological and social benefits it brings. Finally, to establish a judgment and selection model, it is necessary to evaluate the environment of other provinces in China. The selection needs to take Saihanba as an example to establish the geographical location of the ecological reserve.

Keywords

Comprehensive Evaluation Model; Principal Component Analysis; Cluster Analysis.

1. Introduction

Saihanba forest farm has great ecological benefits and has made important contributions to wind resistance and sand fixation and maintaining ecological balance and stability. On the basis of fully collecting data [1], this paper selects appropriate environmental impact assessment indicators, establishes the quantitative environmental impact assessment model of Saihan dam, and compares and evaluates the environmental conditions before and after the restoration of Saihan dam.

Secondly, an impact assessment model is established to analyze the impact of Saihan dam on Beijing sandstorm. Finally, a judgment model is established to determine the area of ecological protection areas to be established in China, and determine the number or scale of ecological protection, so as to further establish the impact evaluation model of time carbon neutralization in China.

2. Principal Component Comprehensive Evaluation Model

Principal component analysis is to delete redundant repeated variables (closely related variables) for all originally proposed variables, and establish as few new variables as possible, so that these new variables are irrelevant, and these new variables maintain the original information as much as possible in reflecting the information of the subject [2].

Use $x_1, x_2, \ldots, x_n$ to represent the selected indicators respectively. Where n is the number of indicators. Use i, m to represent different years and total years, and the index values of year I are recorded as $[a_{i1}, a_{i2}, \ldots, a_{in}]$,the construction matrix is $A = (a_{ij})_{m\times n}$.

Calculate the correlation coefficient matrix $R=(r_{ij})_{n\times n}$.

$$r_{ij} = \frac{\sum_{k=1}^{m} \overline{a_{ki}} \cdot \overline{a_{kj}}}{m-1} \quad (1)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0080

Calculate eigenvalues and eigenvectors. Calculating the eigenvalue of the correlation coefficient matrix R ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$), and the corresponding standardized eigenvector. In the middle, the eigenvector forms a new index variable, that is y_n.

$$\begin{aligned} y_1 &= u_{11}\widetilde{x_1} + u_{21}\widetilde{x_2} + \cdots + u_{n1}\widetilde{x_n}, \\ y_2 &= u_{12}\widetilde{x_1} + u_{22}\widetilde{x_2} + \cdots + u_{n2}\widetilde{x_n}, \\ &\vdots \\ y_n &= u_{1n}\widetilde{x_1} + u_{2n}\widetilde{x_2} + \cdots + u_{nn}\widetilde{x_n}, \end{aligned} \tag{2}$$

Select p principal component and calculate the comprehensive evaluation value.

Calculate the information contribution rate and cumulative contribution rate of the eigenvalue, which are $b_j$and α_p. Generally, the index variable with cumulative contribution rate close to 1 is used as the principal component. Further, the selected principal components are comprehensively analyzed.

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{n} \lambda_k} \tag{3}$$

Calculate the comprehensive score of each principal component.

$$Z = \sum_{j=1}^{p} b_j y_j \tag{4}$$

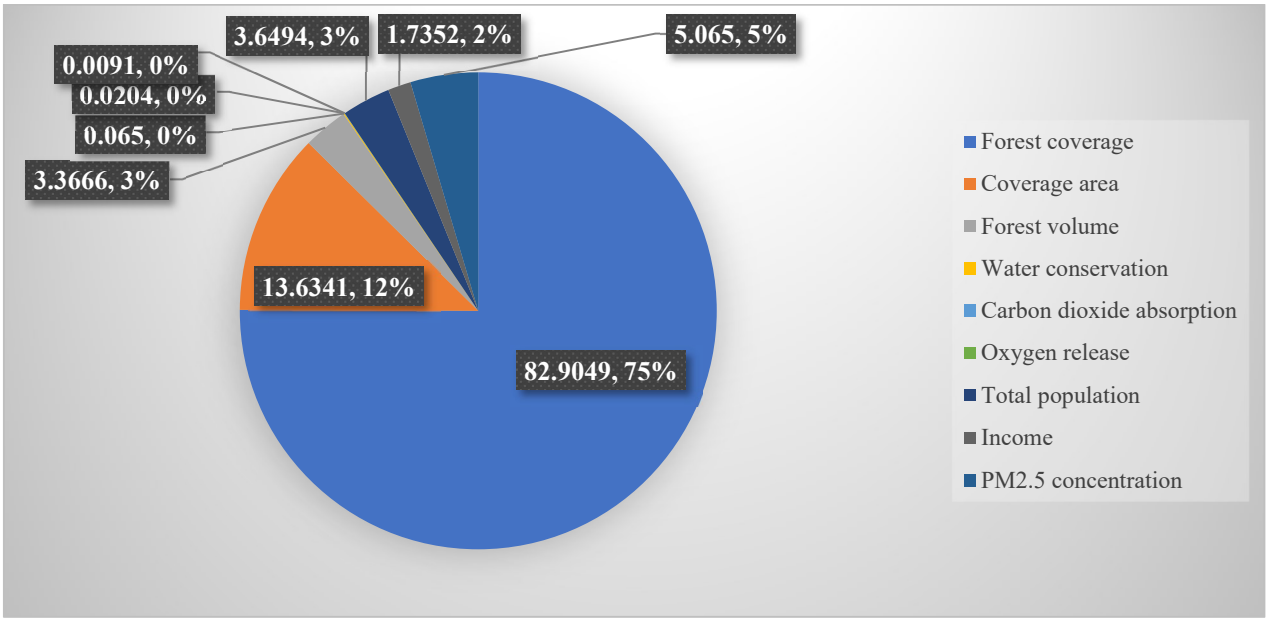

Figure 1. Contribution rate

The cumulative contribution rate of the first three characteristic roots is more than 95%, and the cumulative contribution rate of the first five characteristic roots is 99%. The effect of principal component analysis is very good. The first five principal components are selected for comprehensive evaluation. The eigenvectors corresponding to the first five eigen roots are shown in Table 1 below.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0080

Table 1. Feature vector

	1th eigenvector	2th eigenvector	3th eigenvector	4th eigenvector	5th eigenvector
$\widetilde{x_1}$	0.337676	0.337676	0.355412	0.357948	0.355412
$\widetilde{x_2}$	-0.306710	-0.306710	-0.061020	0.174612	-0.061020
$\widetilde{x_3}$	-0.332900	-0.332900	0.417277	0.075491	0.417277
$\widetilde{x_4}$	0.070452	0.070452	0.164377	-0.909590	0.164337
$\widetilde{x_5}$	0.140790	0.140790	-0.044710	-0.053050	-0.044710
$\widetilde{x_6}$	0.395336	0.395336	-0.012900	0.074287	-0.012900
$\widetilde{x_7}$	-0.286689	0.286688	-0.293358	1.3×10^{-14}	0.741040
$\widetilde{x_8}$	-0.645527	0.645527	0.091339	2.80×10^{-14}	-0.323226
$\widetilde{x_9}$	0.033233	-0.033233	-0.756482	5.49×10^{-14}	0.114239

The principal component comprehensive evaluation model is constructed with the contribution rate of the five principal components as the weight.

$$\mathrm{Z} = 0.8290y_1 + 0.1363y_2 + 0.0337y_3 + 0.0007y_4 + 0.0002y_5 \quad (5)$$

Table 2. Ranking and comprehensive evaluation results

Particular year	Ranking	Evaluation value	Particular year	Ranking	Evaluation value
2019	1	6.210	1990	30	-0.609
2018	2	5.379	1989	31	-0.676
2017	3	4.686	1988	32	-0.759
2016	4	4.114	1987	33	-0.863
2015	5	3.549	1986	34	-0.977
2014	6	3.234	1985	35	-1.090
2013	7	2.968	1984	36	-1.195
2012	8	2.730	1983	37	-1.279
2011	9	2.531	1982	38	-1.352
2010	10	2.323	1981	39	-1.427
2009	11	2.162	1980	40	-1.502
2008	12	1.983	1979	41	-1.579
2007	13	1.894	1978	42	-1.656
2006	14	1.775	1977	43	-1.735
2005	15	1.633	1976	44	-1.814
2004	16	1.482	1975	45	-1.895
2003	17	1.322	1974	46	-1.978
2002	18	1.193	1973	47	-2.060
2001	19	0.955	1972	48	-2.142
2000	20	0.714	1971	49	-2.226
1999	21	0.543	1970	50	-2.311
1998	22	0.384	1969	51	-2.397
1997	23	0.236	1968	52	-2.485
1996	24	0.097	1967	53	-2.566
1995	25	-0.034	1966	54	-2.656
1994	26	-0.159	1965	55	-2.740
1993	27	-0.278	1964	56	-2.833
1992	28	-0.393	1963	57	-2.919
1991	29	-0.506	1962	58	-3.007

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0080

By substituting the five main component values of each year into the above formula, the ranking and comprehensive evaluation results of each year can be obtained, as shown in Table 2 below.

Table 2 shows the comprehensive evaluation and ranking of the ecological environment of Saihanba since 1962. It can be seen from Table 3 that the environmental score of Saihanba has made great progress from negative value to positive value. Through the ranking of years, the comprehensive evaluation score has been increasing. This indicates that in decades, Saihanba has realized the transformation from a very bad natural environment to a forest sea. The restoration of Saihanba has a very positive impact on the ecological environment.

3. Impact Assessment Model

In order to make the expressiveness of each variable the same, it is necessary to standardize the data.

$$\begin{gathered} \overline{a_j} = \frac{\sum_{i=1}^{m} a_{ij}}{m} \\ s_j = \sqrt{\frac{\sum_{i=1}^{m} (a_{ij} - \overline{a_j})^2}{m-1}} \\ b_{ij} = \frac{a_{ij} - \overline{a_j}}{s_j} \quad j = 1,2,\cdots,n \end{gathered} \tag{6}$$

The priority of alternatives in scheme set D can be arranged.

The concept of using the ideal solution to solve the multi-attribute decision-making problem is simple. If an appropriate distance measure is defined in the attribute space, the distance between the alternative solution and the ideal solution can be calculated. TOPSIS method uses Euclidean distance [3]. When only the positive ideal solution is used, sometimes the distance between two alternatives and the positive ideal solution is the same. In order to distinguish the advantages and disadvantages of the two schemes, the negative ideal solution is introduced and the distance between the two schemes and the negative ideal solution is calculated. The scheme with the same distance from the positive ideal solution is far from the negative ideal solution [2].

The standard decision matrix is obtained by vector programming method. Let the decision matrix of multi-attribute decision-making problem be $\mathrm{A} = \left(a_{ij}\right)_{m\times n}$, and the normalized decision matrix be $\mathrm{B} = \left(b_{ij}\right)_{m\times n}$.

$$b_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{m} {a_{ij}}^2}} \tag{7}$$

The weighted gauge matrix is constructed as $\mathrm{C} = \left(c_{ij}\right)_{m\times n}$. Let the weight vector of each attribute given by the decision-maker be $w = [w_1, w_2, \cdots, w_{\mathrm{m}}]^{\mathrm{T}}$.At this time, the weighted gauge matrix is $\mathrm{C} = \left(c_{ij}\right)_{m\times n} = \left(w_{ij} \cdot b_{ij}\right)_{m\times n}$.

Determine the positive ideal solution C^* and the negative ideal solution C^0. Let the jth attribute value of positive ideal solution C^* be c_j^* and the j-th attribute value of negative ideal solution C^0 be c_j^0.

$$\text{Positive ideal solution } c_j^* = \begin{cases} \max c_{\mathrm{ij}} & \text{j is a benefit type attribute} \\ \min c_{\mathrm{ij}} & j \text{ is a cost type attribute} \end{cases} \tag{8}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0080

$$\text{Negative ideal solution } c_j^0 = \begin{cases} minc_{ij} & j \text{ is a benefit type attribute} \\ maxc_{ij} & j \text{ is a cost type attribute} \end{cases} \quad (9)$$

Calculate the distance from each scheme to the positive ideal solution and the negative ideal solution. The distance from alternative d_i to the positive ideal solution is s_i^*.

$$s_i^* = \sqrt{\sum_{j=1}^{n}\left(c_{ij} - c_j^*\right)^2} \qquad i = 1,2,\cdots,m \quad (10)$$

Calculate the ranking index value of each scheme (The comprehensive evaluation index is f_i^*):

$$f_i^* = \frac{s_i^0}{(s_i^0 + s_i^*)} \qquad i = 1,2,\cdots,m \quad (11)$$

4. Judgment Selection Mode

Cluster analysis is to establish a classification method to automatically classify a batch of sample data (or variables) according to their affinity in nature without preconditions [4,5].

Carbon emission and absorption are used to measure the demand of ecosystem for greening. Taking five years as a cycle, the carbon emission of the province numbered i in the j cycle is W_{ij}, the forest area is S_{ij}, and the carbon absorption is Q_{ij}. Here, the average consumption d_m of oil, natural gas and other energy in each province every five years is used to calculate the carbon emission. If the carbon emission coefficient per cubic meter of the energy is t_m, the carbon emission of the province in a cycle is W_{ij}.

$$\mathrm{W_{ij}} = \sum_{\mathrm{m=1}}^{5} \mathrm{d_m} \cdot \mathrm{t_m} \quad (12)$$

Similarly, let the carbon absorption coefficient per hectare of forest be p.

$$\mathrm{Q}_{ij} = \mathrm{p} \cdot S_{ij} \quad (13)$$

Here, the demand index φ_i is defined to represent the demand of different provinces for strengthening greening and implementing the ecological protection mode of Saihan dam. The greater the demand index, the more necessary it is for the province to establish ecological reserves.

$$\delta_{ij} = \frac{\mathrm{W}_{ij}}{\mathrm{Q}_{ij}} \quad (14)$$

$$\varphi_i = \frac{\sum_{j=1}^{M} \delta_{ij}}{M} \quad (15)$$

First, number the collected 30 provinces. The numbering results are shown in Figure 2 below.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0080

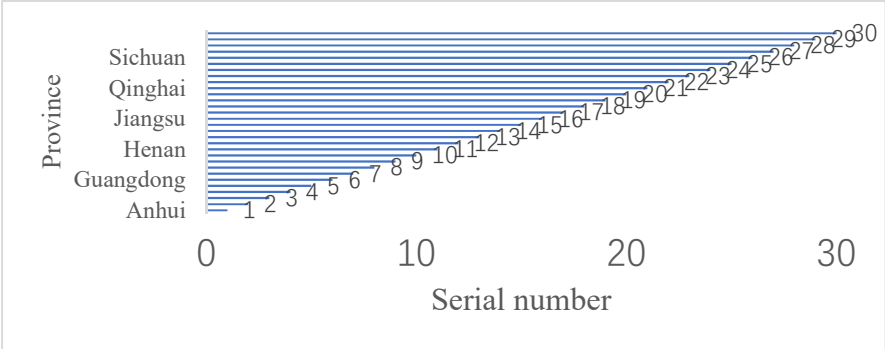

Figure 2. Province number

Based on the collected data, cluster analysis is carried out with MATLAB software, and the results are shown in Figure 3 below.

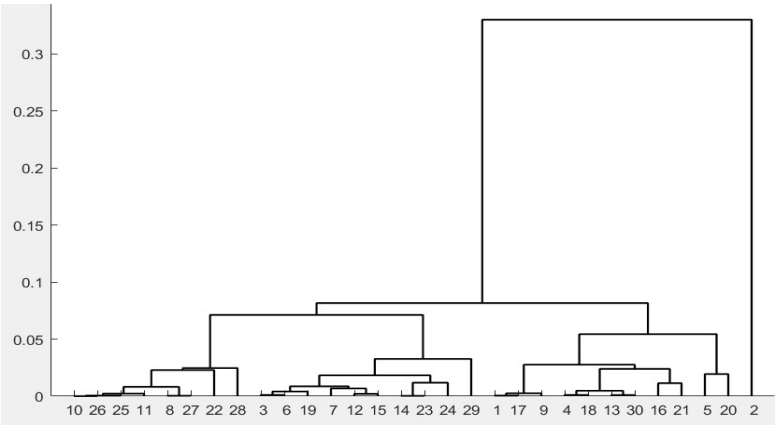

Figure 3. Cluster analysis results

The cluster analysis results are sorted into tables, as shown in Table 3 below.

Table 3. Cluster analysis results

Class serial number	Included sequence number
①	28
②	8 10 11 22 25 26 27
③	1 9 17
④	4 13 16 18 21 30
⑤	29
⑥	3 6 7 12 14 15 19 23 24
⑦	5 20
⑧	2

As can be seen from Table 3, the province most in need of establishing ecological areas in China is Xinjiang, followed by Guizhou, Hebei, Henan etc.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202208_8(8).0080

5. Model Evaluation and Further Discussion

5.1 Strengths

1) In section 2, nine evaluation indexes are fully considered, and the missing data are processed by regression equation to obtain more reasonable ecological environment indexes.

2) By calculating the weight coefficients of relevant influencing factors by principal component analysis, the impact of Saihan dam restoration on the ecological environment can be obtained scientifically and intuitively.

3) Establish the demand index with the ratio of carbon emission to absorption as the medium, which is simple to operate and has clear analysis ideas.

5.2 Weaknesses

1) The actual carbon emission and carbon absorption are not completely consistent with the assumptions, which will affect the accuracy of the model in practical application. In the practical application of the model, the carbon emission and absorption can be corrected according to the actual situation.

2) The indicators that actually reflect the ecological environment are very complex, and the consideration is still not comprehensive enough.

References

[1] https://www.saikr.com/apmcm/2021.

[2] Xiaohai Han, Yaohui Zhang, Fujun Sun, Shaohua Wang. Determination method of index weight based on principal component analysis [J] Journal of Sichuan ordnance industry, 2012,33 (10): 124-126.

[3] Haitian Shi, Fei Deng, Xuepeng Song. Research on blasting safety evaluation based on critical-g1-topsis method [J] Chemical minerals and processing, 2022,51 (04): 41-44 + 50.

[4] Yong Zhouzhang, Wang Yan. Multi attribute bottleneck region recognition method based on interval number cluster analysis [J] Modern manufacturing engineering, 2022 (01): 1-9.

[5] Guanglan Zhou, Fangyong Wu. Research on regional logistics in Anhui Province Based on entropy weight method and cluster analysis [J] Journal of Huainan Normal University, 2022,24 (01): 77-82.