Research on Higher Education Level in Different Regions of China--Based on factors and clustering

Shiwei Wang *
Minzu University of China, Beijing, China, 100081
* Corresponding Author Email: a1779669705@163.com

Abstract. This paper uses factor analysis and cluster analysis to evaluate and classify the development level and status of higher education in various regions of China. It is found that the level of educational development is greatly related to the economic strength and urban development level of various regions, and there is an imbalance between the development of the east and the west in the level of higher education, especially the development level of ethnic minority areas and the weak development level of the western region.In order to promote the comprehensive, coordinated and sustainable development of China's higher education, regional governments should make improvements, such as promoting the construction of new methods of higher education such as MOOC.

Keywords: factor analysis, cluster analysis, development level of higher education.

1. Introduction

Since the Communist Party of China, the educational outlook of our country has undergone structural changes. In 2021, the average number of years of education for the working-age population reached 10.9 years, 1.0 years higher than in 2012. Among them, the proportion of higher education was 24.9%, 10.3 percentage points higher than in 2012. Since the reform and opening-up,the policy and practice of "education equity", which originated from compulsory education and pointed to individual educational opportunities, has been gradually implemented, from policy to system improvement step by step, from goal to gradual clarity, which has laid a solid foundation for promoting social equity and building a harmonious society[1] .There is still a gap between enrolment opportunities in backward rural areas and developed urban areas. At the same time, the support of talents, funds, technology and other resources in first-class colleges is significantly higher than that in ordinary colleges. These phenomena will inevitably widen the gap between higher education levels in different regions and affect the enthusiasm and space of China's educational construction [2] .

Therefore, this paper takes 31 provinces and cities as the evaluation objects, selects the original data of higher education level indicators of all provinces and cities in China in 2020, and studies the current situation of higher education in various regions of China with the help of factors and clustering. The results of this study provide visual results for the current situation of China's higher education level, and at the same time put forward feasible suggestions on how to promote the balance of China's higher education level.

2. Analysis of the level of development of higher education

2.1. Data Sources

This paper takes 31 provinces and cities in China as the evaluation object, refers to the index selection of Zhang Zhiyuan & Hu Shu [3] and combines the relevant statistical yearbooks, and selects the original data of the indicators of higher education development level of various provinces and cities in China in 2020, as shown in Table 1, which is derived from the "2020 China Education Statistics Yearbook", "2020 China Education Expenditure Statistical Yearbook" and "2020 China Statistical Yearbook".

2.1.1 Symbol description

Table 1. Symbol description

Symbol	Description
x1	Total number of students per 100,000 population
x2	Total number of faculty and staff
x3	Number of full-time teachers
x4	Average school building area per student
x5	Number of teachers with senior professional titles
x6	Expenditure on education per student
x7	Graduates per 100,000 population
x8	Number of degrees awarded per 100,000 population
x9	Per capita higher education public finance budget education expenses
x10	Enrollment per 100,000 population
x11	Number of classrooms per student

2.2. Data preprocessing

Due to the large differences in population between provinces and cities, the original data in the Statistical Yearbook of Education are not comparable. This paper combines the corresponding population (unit: 100,000) in each region of the 2020 China Statistical Yearbook to obtain the total number of students per 100,000 population, the number of graduates, the number of enrollment and the number of degrees awarded.

Due to the large differences in the scale of education in different provinces and cities, and the original data on school building area and number of classrooms are not comparable, this paper combines the total number of higher education students in each region in the 2020 China Education Statistics Yearbook to obtain the number of classrooms per student and the area of school buildings per student.

Table 2. Descriptive statistical analysis

Variable	Mean	Median	Minimum	Maximum
x1	3060.35	2866	1499	5393
x2	87134.94	80200	3981	181702
x3	59739.77	52231	2712	133898
x4	47.15	45.08	30.22	72.35
x5	25829.55	23227	1134	62781
x6	3911.45	3322.4	2382.14	8271.75
x7	783.82	797.05	409.78	1401.92
x8	388.7	314.27	181.97	1125.15
x9	25845.89	22160.73	15275.35	88592.53
x10	987.76	983.5	466.04	1534.76
x11	0.02	0.02	0.01	0.03

Due to the different units of the original variables and the descriptive statistical analysis results showing that the numerical variability of each variable is quite different, this paper uses R software to standardize the original variables and then study the data.

2.3. Research methods

2.3.1 Factor analysis

In this paper, factor analysis is used to construct a higher education level evaluation model. Factor analysis is a multivariate statistical method that simplifies and analyzes high-dimensional data. The mathematical model of factor analysis is as follows [5]:

$$\begin{cases} X_1 = \mu_1 + a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + \varepsilon_1 \\ X_2 = \mu_2 + a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + \varepsilon_2 \\ \cdots \\ X_p = \mu_p + a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + \varepsilon_p \end{cases} \tag{1}$$

Where $X = (X_1, X_2 \dots X_p)'$ is an observable p–dimensional random vector with a mean:$E(X) = \mu$ and a covariance matrix:$Cov(X) = \Sigma. \mu = (\mu_1 \dots \mu_p)'$ is the mean vector of X,$f = (f_1 \dots f_m)'(m < p)$ is an unobservable random vector with $E(f) = 0, Cov(f) = I_m, \varepsilon = (\varepsilon_1 \dots \varepsilon_p)', E(\varepsilon) = 0, Cov(\varepsilon) = D = diag(\sigma_1^2 \dots \sigma_p^2), Cov(f, \varepsilon) = 0$,and matrix A is the factor loading matrix.

Therefore, (1) can be expressed as $X = \mu + Af + \varepsilon$. Af Represents the part that can be explained by m principal components, and ε represents the remainder with a small amount of information.

2.3.2 Cluster analysis

The fundamental principle of clustering analysis is to group together more comparable objects while separating individuals with significant dissimilarities into distinct clusters. The hierarchical clustering function of the R language is a fundamental feature, providing statistical calculation and random number generation functions. Hierarchical clustering algorithms do not need to predetermine the number of clusters, and the similarity of distances and rules is easy to define and less restrictive, and arbitrary shapes can be found [6]. In this paper, the relationship between samples is measured by defining distances by the sum of squared scattering methods.

The sum of the squared Euclidean distances from each sample in the class to the center of gravity of the class is called the sum of squared (within-class) dispersions.Let class the G_K and G_L merge into a new class G_M,then the sum of squared dispersions of G_K, G_L, G_M are:

$$W_K = \sum_{i \in G_K} (x_i - \overline{x_K})'(x_i - \overline{x_K}), W_L = \sum_{i \in G_L} (x_i - \overline{x_L})'(x_i - \overline{x_L}), W_M = \sum_{i \in G_M} (x_i - \overline{x_M})'(x_i - \overline{x_M})$$

For a fixed number of in-class samples, they reflect the degree of dispersion of the samples within the respective class. If the categories G_K and $G_L$are close together, the sum of squared dispersions is added after merging. $W_M - W_K - W_L$ should be small; otherwise, it should be larger. So we define the square distance between G_K and G_L as [7]:

$$D_{KL}^2 = W_M - W_K - W_L \tag{2}$$

The sum of squares method of dispersion makes it difficult to merge two large classes.On the contrary, two small classes are easy to merge, which often meet our practical needs for cluster analysis.

3. Analysis of results

3.1. Analysis of factor analysis results

3.1.1 Correlation test

The underlying assumption of factor analysis is that there is a high degree of correlation among observed variables. To assess the suitability of applying factor analysis to the given set of variables, the KMO and Bartlett spherical tests are performed. The results indicate that the KMO value is 0.85, which is above the recommended minimum value of 0.7. In addition, Bartlett spherical tests provides a statistical significance level of $p < 0.001$, indicating that the data show a strong correlation and is suitable for factor analysis.

Table 3. KMO and Bartlett spherical tests

KMO measurement		0.85
Bartlett's sphericity test	Approximate chi-square	512.874
	df	55
	sig	0.000

3.1.2 Extraction of factors

In this paper, a common factor is extracted by spindle iterative method, followed by factor rotation. The number of common factors extracted is then determined based on the cumulative contribution rate. Using two factors, the cumulative contribution rate reaches 68%. After extracting three factors, the cumulative contribution rate amounts to 84%. Using four common factors, the cumulative contribution rate rose to 92 per cent. This represents a substantial amount of information, and the cumulative contribution rate is still marginally increasing with the continued extraction of common factors. Therefore, the first four common factors are chosen for factor analysis.

Table 4. Factor cumulative contribution rate

Factor	1	2	3	4
Percentage variance	0.33	0.3	0.18	0.11
Cumulative contribution rate	0.33	0.63	0.81	0.92

3.1.3 Factor rotation

The factor load value represents the degree of association between the main factor and the original variable, and a larger factor load value represents more metric information about the original variable that the factor contains. In order to rationally explain the common factors, the maximum variance rotation method is selected for factor rotation, and finally the factor loading matrix of the four main factors is obtained as shown in Table 4.

Table 5. Rotation factor load matrix

Symbol	Fator			
	1	2	3	4
x1	0.92	0.3	-0.01	-0.1
x2	0.24	0.96	-0.08	-0.16
x3	0.13	0.95	-0.18	-0.18
x4	-0.18	-0.52	0.22	0.85
x5	0.31	0.94	-0.02	-0.13
x6	0.18	-0.06	0.96	0.02
x7	0.97	0.23	-0.04	-0.08
x8	0.93	0.04	0.22	0.04
x9	-0.12	-0.18	0.92	-0.08
x10	0.89	0.31	-0.13	-0.19
x11	-0.1	-0.3	-0.13	0.94

The loading matrix for the rotation factor indicates that the initial primary factor bears more weight on the original variables x1, x7, x8, and x10, with respective values of 0.92, 0.97, 0.93, and 0.89. Therefore, the first primary factor can be inferred as the indicator for the scale of development of higher education. The updated text maintains a precise, formal, and proficient style by implementing appropriate terminology and phrasing.

The second primary factor exhibits elevated loadings on x2, x3, and x5, with values of 0.96, 0.95, and 0.94, respectively. Thus, it can be interpreted as the input of human resources into higher education with greater precision and formality.

The third most significant factor exhibits a greater load on variables x6 and x9, namely 0.96 and 0.92, respectively, thus indicating that it can be identified as funding for education.

The fourth most significant factor exhibits higher loading values on variables x4 and x11, which are 0.85 and 0.94, respectively. Therefore, it can be interpreted as the input of material resources into higher education.

3.1.4 Build factor score models

Combined with the cumulative contribution rate and factor scores of each region, this paper constructs a comprehensive evaluation model of higher education according to formula (3) through the literature research method and the factor score model constructed by Wang Shuo and Ma Lijun [4].

$$F_C = \sum_{i=1}^{4} F_i u_i / \sum_{i=1}^{4} u_i \tag{3}$$

$$\text{Derived:} F_C = \frac{0.33F_1+0.30F_2+0.18F_3+0.11F_4}{0.92} \tag{4}$$

Using factor scores from 31 provinces and cities in formula (4) of the comprehensive assessment model, the overall score for the level of development of higher education for each region can be obtained, accurate to three decimal places. In addition, the corresponding global rankings can be found in Table 6.

Table 6. Scores and rankings of the comprehensive development Level of higher education in each region

Region	Scale of education	Faculty	Funding	Material investment	comprehensive	
	Score	Score	Score	Score	Score	Rank
Beijing	2.991	1.824	2.236	0.319	1.971	1
Tianjin	0.293	1.877	0.201	0.995	0.876	2
Hebei	1.313	-0.117	1.699	0.718	0.689	3
Shanxi	1.197	0.290	-0.104	0.521	0.566	4
Inner Mongolia	0.559	0.838	0.043	-0.016	0.480	5
Liaoning	-0.797	1.963	0.990	-1.714	0.343	6
Jilin	2.511	-1.493	-0.149	-0.383	0.339	7
Heilongjiang	-0.314	1.731	-0.336	-0.429	0.335	8
Shanghai	-0.443	0.909	-0.840	0.121	0.286	9
Jiangsu	-0.667	0.686	0.808	0.351	0.185	10
Zhejiang	-0.707	1.229	-0.267	-0.007	0.094	11
Anhui	0.466	0.053	-0.597	0.131	0.083	12
Fujian	0.047	-0.077	-0.389	1.053	0.042	13
Jiangxi	0.195	0.412	-0.594	-0.462	0.033	14
Shandong	0.263	-0.034	-0.669	-0.350	-0.090	15
Henan	-0.045	-1.066	-0.085	2.320	-0.103	16
Hubei	-0.488	-0.274	-0.513	1.873	-0.141	17
Hunan	-0.386	0.542	-0.687	-0.687	-0.178	18
Guangdong	0.861	-1.134	-0.638	-1.403	-0.191	19
Guangxi	-0.731	-0.370	-0.264	1.803	-0.219	20
Hainan	-0.216	-0.369	-0.815	0.859	-0.255	21
Chongqing	0.394	-0.701	-0.544	-0.591	-0.264	22
Sichuan	-0.574	-0.223	-0.033	-0.652	-0.363	23
Guizhou	-0.519	-0.066	-0.620	-0.676	-0.410	24
Yunnan	0.031	-0.286	-0.680	-1.849	-0.436	25
Tibet	-1.599	-1.234	3.304	-1.049	-0.455	26
Shaanxi	-1.440	-1.009	1.492	0.562	-0.486	27
Gansu	-0.998	-0.626	-0.151	0.540	-0.527	28
Qinghai	-0.468	-0.659	-1.036	-0.310	-0.623	29
Ningxia	-0.348	-1.048	-0.464	-0.558	-0.624	30
Xinjiang	-0.381	-1.569	-0.297	-1.030	-0.829	31

3.2. Cluster analysis results

Based on the results of factor analysis, this paper uses the sum of squared dispersion method to cluster the normalized factor scores of 31 provinces and cities,

The classification results are shown in the following table：

Table 7. Cluster analysis results

Category	Province and city
The first category	Beijing
The second category	Tianjin, Jiangsu, Hubei, Shaanxi,Shanghai
The third category	Guangdong, Zhejiang, Sichuan, Shandong, Henan
Category IV	Liaoning, Jiangxi, Hunan, Hebei, Anhui, Fujian, Hainan, Yunnan, Gansu, Guangxi, Jilin, Chongqing
Category V	Ningxia, Inner Mongolia, Guizhou, Xinjiang, Shanxi, Heilongjiang
Category VI	Tibet ,Qinghai

These regions are stratified into six tiers, from first to sixth, and have experienced a decrease in the level of tertiary education.

We classify the level of higher education in 31 provinces and cities across the nation into four tiers.

The highest tier of higher education in China comprises Beijing, Tianjin, Shanghai, and six provinces, with Beijing boasting the most distinguished level of higher education and being classified separately. Beijing holds a relatively prominent position in areas such as education, human capital, and economic investment. The other five regions lead in the composite factor score, but their development is uneven, and Shanghai, Tianjin, and Shaanxi need to increase investment in educators.

The second tier is made up of five provinces, such as Guangdong and Zhejiang, except for Sichuan, which are located in central or eastern regions. Though the level of higher education development in these regions may be lower than in the first echelon, it still ranks at the forefront of educational development. These regions benefit from higher quality educators. However, their level of education lags behind that of the premier provinces.

The third echelon consists of Liaoning, Jiangxi, and another 12 provinces. The educational standard of these provinces is above and below the average level of higher education in China. These provinces have their own strengths, but the benefits are weaker than the drawbacks. The level of investment in both capital and human resources is inadequate.

The fourth echelon is comprised of Ningxia, Inner Mongolia, and another 8 provinces. Most of these provinces are located in the western region of China and are located in ethnic minority areas and economically underdeveloped regions of China. These regions have relatively sluggish economic development and low levels of higher education. At the same time, Tibet and Qinghai are classified separately, in conjunction with the findings of factor analysis, because although their overall education level is backward, they receive a high score in funding. This is closely related to the state's policy of care for these two regions, as well as their population size.

4. Conclusions and recommendations

4.1. Conclusion

In terms of the scale of education, Beijing, Tianjin, and Shanghai rank among the top three. This is closely linked to robust economic growth in these three regions and the profusion of accessible universities. Aspiring students aim to enroll in prestigious educational institutions located in metropolitan areas, with the goal of expanding their horizons and improving their skills.

In terms of educators, Guangdong, Jiangsu, and Beijing ranked in the top three. These three areas boast favourable working conditions, ample opportunities and resources, and improved remuneration. Excellent teachers are often willing to choose such urban developments, so the overall quality of teachers in these areas is relatively high. Teachers in Tianjin, Hainan, Tibet and Jilin are relatively weak.

In terms of funding, Tibet, Beijing and Shanghai ranked among the top three. As first-tier cities in China, Beijing and Shanghai have good economic development, more capital investment, and a large scale of educational development, which is in line with our actual understanding. However, due to the small number of students in Tibet and the large per- student funding, at the same time, during the "Thirteenth Five-Year Plan" period, the information construction of Tibet's colleges and universities focused on functional requirements such as technology popularization and business capabilities, ensuring supply, and reconstruction and application[8]. The state has established a "state-based, provincial-level" Tibet education funding guarantee mechanism for Tibet, so Tibet, as a region with a low overall score, is in a high position in terms of funding.

In terms of material investment, Guizhou, Ningxia and Inner Mongolia scored higher, which is different from our expected results. According to the analysis, since the indicators selected in this data analysis in terms of material investment are material investment per student, and the urban resources and space in developed areas such as Beijing, Shanghai and Jiangsu are limited and the number of students is large, it is difficult to make breakthroughs in material investment. Ningxia, Guizhou and Inner Mongolia are sparsely populated, and they score very high in material investment in higher education.

On the whole, China's higher education level is still in a very uneven state, among which Beijing's education level is much higher than that of other cities; provinces and cities such as Shanghai, Jiangsu, Guangdong, Tianjin, Hubei and Shaanxi also have higher education levels. Higher education is greatly affected by regional development, and most of the regions with the lowest levels of educational development are marginal provinces. Differences in economic development have led to an imbalance in human resources and a widening gap in education levels. Most regions have both advantages and shortcomings in the level of education, while there are fewer regions with higher score levels of various factors, and the gap between the level of higher education in the east and west of China is large, how to promote the balanced development of higher education in various regions in China and enhance the competitiveness of colleges and universities in the western region is an urgent problem to be solved.

4.2. Suggestion

(1)The government should increase funding for higher education in economically underdeveloped areas.

Through the results of factor analysis and cluster analysis, we can see that the level of development of higher education in economically developed areas is also high, and the level of higher education in economically weak western regions and ethnic minority areas is also very weak, so the State should continue to increase investment in education and the treatment of teachers in economically weak areas, increase funding and material investment, especially in areas where the scale of education is larger but materials and funds are relatively weak, only to continuously increase economic investment in such areas. To reduce imbalances in higher education and promote equity in education.

(2)Pay attention to the balanced development of the quality of education.

First of all, from the macro level: the relevant education regulations and policies formulated by governments at all levels and education departments should reflect the basic idea of balanced development of education; In the process of promoting the development of higher education, it is necessary to promote the idea of balanced development of higher education to guide education work [9].

Secondly, each region should combine the current situation, consolidate the advantages, make up for the shortcomings, improve the comprehensive strength of the education level, and only by improving the comprehensive strength can we attract talents, and then use excellent educational resources to cultivate talents, expand the scale of education, enhance the competitiveness of regional education, and then promote the balanced development of education quality.

(3) Promote the construction of new methods of higher education such as MOOC

In the era of rapid development of digitalization, more and more emerging methods of education have flooded the market. For example, MOOCs and other online course teaching platforms. MOOC can enable more people to receive higher education, even high-quality higher education[10]. Therefore, the state should actively build network information construction in various regions, increase the investment of material resources, and alleviate the imbalance of China's higher education level in the short term when teacher resources cannot be balanced.

References

[1] Wu Daguang, Wang Yiqian. Some Higher Education Levels in Eastern and Western China [J]. Journal of Lanzhou University (Social Science Edition), 2021, 49 (05): 1-8. DOI: 10.13885/j.issn.1000-2804.2021.05.001.

[2] Hu Yu. An Empirical Study on Regional Differences of Higher Education Level--Based on SPSS Factor [J]. Jiangsu Higher Education, 2019 (06)

[3] Zhang Zhiyuan, Hu Shu. Research on Inter-provincial Differences of Higher Education Level in China--Based on Education Statistics in 2018 [J]. Shandong Higher Education, 2021, 9 (02): 38-46.

[4] Wang Shuo & Ma Lijun. Research on the Differences of Regional Higher Education Level in China--Based on the Data of China Statistical Yearbook in 2016 [J]. Journal of Xinyang Normal University (Philosophy and Social Sciences Edition), 2019, 39 (03): 49-54.

[5] Li Haiyan. Evaluation of Listing Profit Based on Factor Method--Taking Anhui Listing as an Example [J]. Friends of Accounting, 2012 (20): 71-74.

[6] Li Mingchan, Kun Yang, Xu Quanli, Zhang Shaohua. Clustering of temporal and spatial distribution of Ludian M_S6. 5 earthquake in 2014 [J]. Geography and Geographic Information Science, 2018, 34 (02): 66-72 +1.

[7] Zhu Jianping. Applied Multivariate Statistics, 4th Edition [M]. Beijing: Science Press, 2021

[8] Li Jing, Zhang Liguo. New educational infrastructure promotes the modernization of higher education: mission direction and challenge [J]. Journal of Minzu University (Philosophy and Social Sciences Edition), 2022, 43 (06): 109-116.

[9] Zhai Bo. Educational Equilibrium: Theoretical Index and Measurement [J]. Educational Research, 2006 (03): 16-28.

[10] Wang Wenli. MOOC and its influence on higher education [J]. Jiangsu Higher Education, 2013 (02): 53-57. DOI: 10.13236/j.cnki.jshe.2013.02.009.