Frontiers in Economics and Management Volume 4 Issue 12, 2023
ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

Research on Influencing Factors of Grain Yield: Evidence based on Provincial Data in China

Chenyuke Jin

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu 233000, China

Abstract

Grain production level is closely related to economic development and social stability in one region. In this paper, multiple linear regression, principal component analysis and cluster analysis are used to explore the main factors affecting grain output in 31 provinces of our country. The 31 regions in China are divided according to the grain production level by cluster analysis, nine agricultural indicators are determined, and the regions are ranked by the principal component analysis method and the grey correlation analysis method. The results show that: China's grain output has the characteristics of regional distribution, which can be divided into three categories; Grain output is not the only factor to measure the level of grain production, but also need to consider the planting area and mechanization degree.

Keywords

Grain Yield; Cluster Analysis; Principal Component Analysis; Grey Correlation Analysis.

1. Introduction

The Chinese nation is one of the oldest farming nations in the world, and the Chinese civilization takes farming civilization as the mother and the main body of its evolution. Even when other industries are rising together in today's society, China is still the world's largest grain producer and the most important exporter of agricultural products. In this regard, it is necessary to understand what factors affect the development of agricultural production in China.

In this paper, the planting industry in agriculture is explored, and the growth of grain output and its influencing factors are taken as the main research direction. In this paper, it is considered that the difference of grain output in different years in the same region is more susceptible to the influence of climate factors and the increasing mechanization level in China.

In order to eliminate the influence of regional differences and single methods on the experimental results, on this basis, a variety of methods were used to analyze the results in several regions of the country. Various agricultural data of more than 30 regions in the country in 2021 were found in the National Bureau of Statistics, and analyzed by cluster analysis, linear regression, principal component analysis and grey correlation analysis.

2. Literature Review

At present, the domestic research on grain yield mainly focuses on multi-factor research and regional research. Multi-factor research, that is, to study the influencing factors of grain yield; Regional research, that is, the study area is a single province and city, so there are many regional limitations.

Among them, research results related to multi factor research include Liping Wu and Jiaxiang Lin [1] In the Analysis of Agricultural Economic Development Level in Main Grain Production Areas by Integrating Principal Component Analysis and Clustering, 17 indicators were selected for principal component and cluster analysis, including the proportion of gross domestic

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

product in the primary industry area, per capita grain yield, urbanization rate, effective irrigation area, agricultural fertilizer use, and grain crop planting area. Hongjing Shi [2] selected 8 main indicators for research in Factors Influencing Grain Yield Level and Cluster Analysis, including total power of agricultural machinery, effective irrigation area, application amount of agricultural fertilizers, rural hydropower stations, and sowing area and affected area of grain crops.

As for the study on regionality, Wenfu Peng [3] used grey Analysis of Factors Affecting Grain Production in Sichuan Province and Grain Output Forecast, in which he analyzed the factors affecting grain output in Sichuan Province from 1994 to 2000 by using grey system analysis. Chunhui Wang, Shenglu Zhou, Shaohua Wu[4] studied the current situation of grain output fluctuation in Jiangsu Province in recent years in their paper Jiangsu Grain Output Forecast Based on Multiple Linear Regression Model and Grey Correlation Analysis. Based on grey correlation analysis, multiple linear regression equation was established to forecast and verify Jiangsu grain output from 2010 to 2009. Finally, relevant suggestions were put forward. Shan Lin[5] made a series of analysis on the grain output in Shaanxi Province in the Analysis of Influencing Factors of grain output. Pengling Liu, Wenjun Wu, Yingying Wan[6] conducted principal component analysis on the grain yield in Anhui Province in their Analysis of Factors Influencing Grain Yield and Grey Prediction - Based on Data from Main Production Areas in Anhui Province, and used a grey prediction model to predict the yield for the next eight years.

Based on the above literature, this article has decided to select effective irrigation area, agricultural fertilizer application amount, total sown area of crops, number of combine harvesters, number of agricultural medium and large tractors, disaster area, disaster rate, disaster area and disaster rate to judge their impact on grain production, among which the disaster rate and disaster rate are calculated according to the disaster area and disaster area. In the following analysis, the disaster area and disaster area are replaced by the two indicators of disaster area and disaster area.

3. Research Methods and Data Sources

With the development of science and technology and the progress of society, the factors affecting food production have gradually become more complex and diversified. When predicting grain output, how to distinguish and identify the influence of various factors, accurately analyze these factors, and clarify and rationalize their impact ranking is the focus of correlation analysis on the influencing factors of grain output. At present, research mainly uses methods such as grey correlation analysis [7] (GRA) and principal component analysis [8] (PCA) to analyze the correlation of factors affecting grain yield. However, further research is still needed to select accurate and effective models to improve the accuracy of grain production prediction.

3.1. Cluster Analysis

Cluster analysis defines the distance between samples and the similarity coefficient between variables, where the distance or similarity coefficient represents the similarity between samples or variables.

In cluster analysis, commonly used distances are Euclidean distance, absolute distance, Mahalanobis distance, etc. In this paper, the Euclidean distance is used to measure the distance between points. The advantage of Euclidean distance is that when the coordinate axis is rotated, the Euclidean distance remains unchanged. Therefore, remember the distance between two points as:

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - y_{jk})^2}$$

Choose to use the longest distance method to measure the distance between classes, and use this method for cluster analysis. At the beginning, the n samples are taken as a class, and the distance between the samples and the distance between the classes are specified. Then the two classes with the closest distance are merged into a new class, and the distance between the new class and other classes is calculated. Repeat the merging of the two nearest classes, reducing one class at a time, until all the samples are merged into one class.

There are two sample classes G1 and G2, where $D(G_1, G_2)$ represents the distance between sample xi belonging to G1 and sample yi belonging to G2. The definition of the longest distance method is:

$$D(G_1, G_2) = \max_{x_i \in G_1, y_i \in G_2} \{d(x_i, y_j)\}$$

3.2. Multiple Linear Regression

Assuming a linear relationship between variable Y and variable $X_1, X_2, \cdots, X_P$:

$$Y = b_0 + b_1 X_1 + \cdots + b_p X_P + \varepsilon, \ \varepsilon \sim N(0, \sigma^2)$$

If $(x_{i1}, x_{i2}, \cdots, x_{i6}, y_i), (i = 1, 2, \cdots n)$ is a set of $n(n > p+1)$ independent predictive values of $(X_1, X_2, \cdots X_P, Y)$, then the multiple linear regression model can be represented as:

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \cdots n$$

Among them, each ε_i is independent of each other.

The following describes the multiple linear regression model in matrix form.

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_3 \end{pmatrix}, b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The above equation can be expressed as:

$$Y = Xb + \varepsilon$$

3.3. Principal Component Analysis

Principal component analysis is the process of reducing multiple variables into a few principal components using dimensionality reduction techniques. These principal components retain most of the information of the original variable, and are often represented as linear combinations of the original variables. Mathematical processing is simply a linear combination of the original p indices into the new indices. The first linear combination, that is, the first

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

comprehensive indicator is denoted as y_1. In order to make the linear combination unique, it is required that the variance of y_1 in all linear combinations be the largest, that is, var(y_1) is the largest, and it contains the most information.If the first principal component is not sufficient to represent all the information of the original p indicators, then consider selecting the second principal component y_2 and require that the existing information of y_1 does not appear in y_2, i.e. cov (y_1, y_2)=0.

Through principal component analysis, a small number of variables can be used to explain most of the changes in the original data, and the highly correlated variables can be transformed into independent or uncorrelated variables, thus achieving the effect of dimensionality reduction. The main component can be identified from these indicators that affect food production, so that the positive influencing factors can be effectively used. If the negative factors also show a significant impact, then we should reduce the harm brought by these factors in agricultural production as much as possible.

3.4. Grey Correlation Analysis

Grey correlation analysis is a systematic theoretical analysis method that uses the order of grey correlation degree to describe the strength, size and order of the relationship between factors. The basic idea of this analysis method is that the higher the similarity or consistency between two variables, the higher the correlation degree between them, and vice versa, the lower the correlation degree between them.

Specific steps of grey correlation analysis:

①Determine the comparison objects (31 regions) and reference series (evaluation criteria). There are 31 experimental data evaluation objects and 7 evaluation indicators, respectively, the effective irrigation area, the conversion amount of agricultural fertilizer application, the total sown area of crops, the number of combine harvesters, the number of large and medium-sized agricultural tractors, the disaster rate and the disaster rate, the reference sequence is $x_0 = \{x_0(k) | k = 1,2,...,n\}$, and the comparison sequence is $x_i = \{x_i(k) | k = 1,2,...,n\}, i = 1,2,...,m$.

② Determine the corresponding weights for each indicator. The weight $w = [w_1, w_2, ..., w_n]$ corresponding to each indicator can be determined using Analytic Hierarchy Process, where $w_k (k = 1,2,..,n)$ is the weight corresponding to the k-th evaluation indicator.

③Calculate the grey correlation coefficient:

$$\xi_i(k) = \frac{\min\limits_s \min\limits_t |x_0(t) - x_s(t)| + \rho \max\limits_s \max\limits_t |x_0(t) - x_s(t)|}{|x_0(k) - x_i(k)| + \rho \max\limits_s \max\limits_t |x_0(t) - x_s(t)|}$$

To compare the correlation coefficient between sequence xi and reference sequence x0 on the k-th indicator, where $\rho \in (0, \infty)$ is the resolution coefficient. Generally speaking, the larger the resolution coefficient ρ, the greater the resolution; The smaller ρ, the smaller the resolution. The value range of ρ is generally (0,1). When $\rho \le 0.5463$, the resolution is best, usually taking ρ=0.5.

④Calculate the gray weighted correlation degree. The calculation formula of grey weighted correlation degree is:

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

$$r_i = \frac{1}{n}\sum_{k=1}^{n} w_i \xi_i(k), k = 1, 2, \ldots, n$$

In the formula, r_i is the grey weighted correlation degree between the i-th evaluation object and the ideal object. This article takes equal weights when calculating the correlation degree.

⑤Evaluation analysis. According to the size of the grey weighted correlation degree, each region is sorted to establish the correlation order, the greater the correlation degree, the higher the grain yield.

3.5. Variable Selection and Data Source

This article comes from agricultural data from the National Bureau of Statistics in 2021, and selects nine common agricultural indicators: effective irrigation area (x1), agricultural fertilizer application amount (x2), total crop planting area (x3), number of combine harvesters (x4), number of agricultural large and medium-sized tractors (x5), disaster area (x6), disaster rate (x7), disaster area (x8), and disaster rate (x9). Each independent variable indicator is set to x1-x9, respectively, The dependent variable indicator grain yield is set to y.

Since it is found that x3(total sown area of crops) in the data is correlated with x6(affected area of crops) and x8(affected area of crops), the final data used for analysis is the ratio of x6, x8 and x3. From the table below, it can be easily seen that areas with larger crop sown area also have larger areas of disaster and disaster. From a practical point of view, we can easily understand this problem, because the greater the total area indicates the greater the likelihood of disaster, and the two are almost necessarily related. Therefore, it is finally decided to use the two indicators x7(crop disaster rate) and x9(crop disaster rate) to replace the disaster area and disaster area for analysis.

4. Empirical Results and Analysis

4.1. Descriptive Statistics of Data

In order to explore the correlation between each influencing factor and the explained variable, a heat map was drawn. The correlation between the dependent variable and each independent variable can be seen according to the size of the circle in Figure 1.

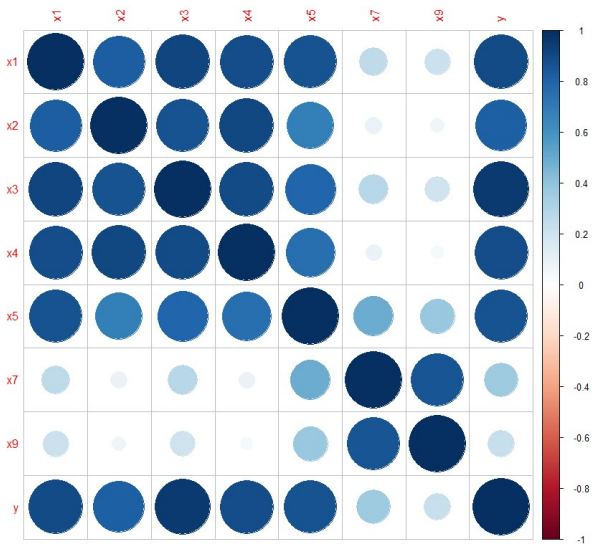

Figure 1. Correlation diagram of each influencing factor

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

Figure 1 shows that the main factors affecting y(grain yield) are x1, x2, x3, x4, x5(effective irrigated area, amount of agricultural fertilizer applied, total sown area of crops, number of combine harvesters, number of large and medium-sized agricultural tractors).

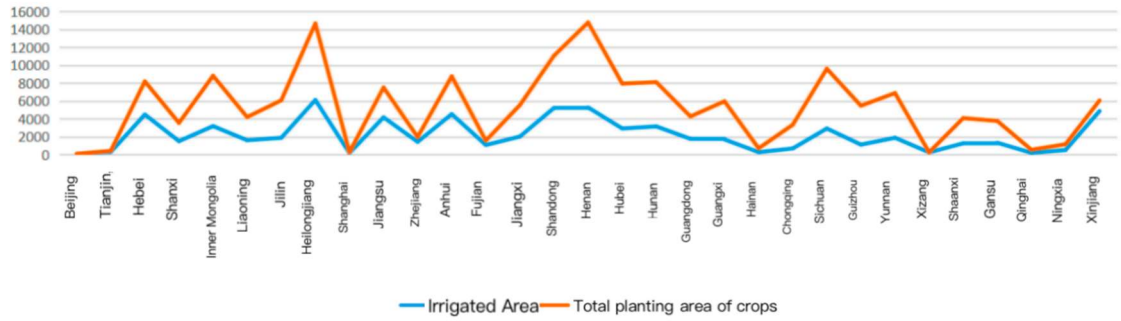

Figure 2. Correlation graph between effective irrigated area and total sown area of crops

It can be seen from Figure 2 that some provinces and autonomous regions have a high total sown area of crops, but the effective irrigated area is less than half of the total sown area, such as Heilongjiang, Henan and Shandong. From the figure, we can see that these three provinces are major grain producing provinces, but their effective irrigated area accounts for a small proportion. This shows that if these regions carry out better planning and management of agricultural water use and develop better irrigation techniques, the grain production in these regions can be further increased. On the contrary, some provinces and cities have a small total sown area of crops, but they can achieve nearly comprehensive effective irrigation, such as Xinjian. Although the grain output is not outstanding, Xinjiang leads the country in the proportion of effective irrigated area. This shows that the application of irrigation technology in Xinjiang is very successful, and also reflects the importance of irrigation technology to agricultural production.

4.2. Results of Cluster Analysis

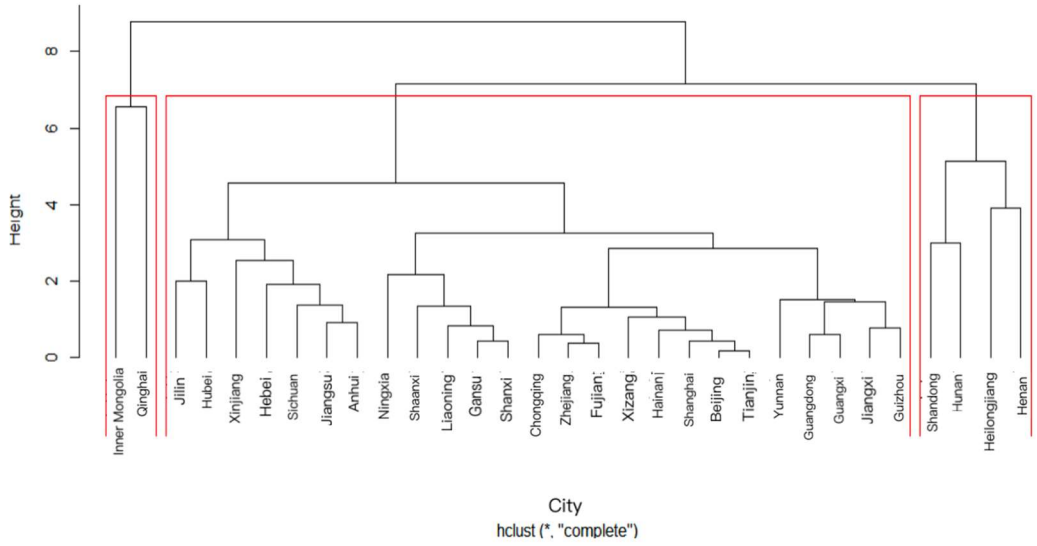

Figure 3. Cluster diagram of the longest distance method

In this paper, the grain output of 31 regions in 2021 was selected for analysis. Seven indicators, including effective irrigation area, agricultural fertilizer application amount, total sown area of crops, number of combine harvesters, number of large and medium-sized agricultural tractors,

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

disaster and disaster rate, were used as variables to carry out cluster analysis for each region. The method of systematic clustering is mainly adopted, the method of European distance in Q-type clustering is used, and the longest distance method is selected to cluster the 31 provinces and cities. The clustering graph is as follows.

From the tree diagram above, the 31 cities can be divided into three categories, as shown in Table 1 below:

Table 1. Three categories of regions

Categories	Regions	Number of districts
First Category	Shandong, Hunan, Heilongjiang, Henan	4
Second Category	Inner Mongolia and Qinghai	2
Third Category	Jilin, Hubei, Xinjiang, Hebei, Sichuan, Jiangsu, Anhui, Ningxia, Shaanxi, Liaoning, Gansu, Shanxi, Chongqing, Zhejiang, Fujian, Xizang, Hainan, Shanghai, Beijing, Tianjin, Yunnan, Guangdong, Guangxi, Jiangxi, Guizhou	25

As can be seen from the table, Shandong, Hunan, Heilongjiang and Henan can be divided into the first category, Inner Mongolia and Qinghai into the second category, and other regions into the third category. Specific analysis is carried out for the first and second categories:

(1) For the first category, Shandong, Hunan, Heilongjiang and Henan are all famous agricultural provinces. Shandong, Hunan and Henan are located in East and Central China. They have mild climate, four distinct seasons, abundant sunshine, abundant rain, fertile land and two crops a year. Located in the northeast border of China, Heilongjiang has a vast territory, a large sown area, and agricultural production is close to the standard of modern agriculture. At the same time, the emphasis on grain production in these areas is relatively high; Therefore, the four regions have a higher level of grain production and can be classified into the same category, which is also consistent with the reality.

(2) Inner Mongolia and Qinghai have similar climate conditions and frequent extreme weather. Affected by frost, drought and other natural disasters, the disaster rate index of Inner Mongolia and Qinghai is relatively high, resulting in low grain production; At the same time, planting and animal husbandry in Inner Mongolia and Qinghai developed together, resulting in a relative decrease in grain production; So it is practical for Inner Mongolia and Qinghai to be divided into one category.

The above analysis shows that agricultural production is significantly affected by region and climate when other conditions are not different. The four regions mentioned in the first category are mostly plain areas with vast land areas suitable for large-scale agricultural production and the use of large-scale agricultural machinery, and naturally produce more grain than other hilly and plateau regions. On the other hand, Qinghai and other regions are inland and have dry climates. In addition, natural disasters occur more frequently, so grain production will naturally decline.

The similarity of agricultural conditions in other categories is also reflected. For example, Beijing, Shanghai, Tianjin and other autonomous regions are grouped together. These regions are economically developed and have a large population density, but their grain output is very small, because these regions are small and their development does not depend on agriculture, and other industries account for a larger proportion of the economy.

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

4.3. Results of Linear Regression

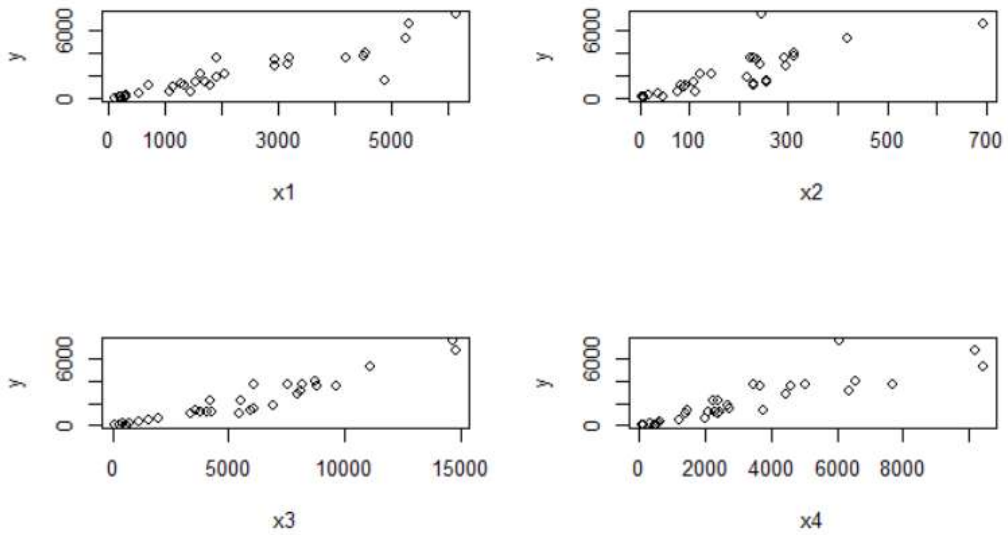

Figure 4. Scatter plot of distribution

The above two plots are the scatter plots of the effective irrigated area, the amount of agricultural fertilizer applied, the total sown area of crops, the number of combine harvesters and the grain output respectively. It can be seen that there is a relatively significant linear relationship. In addition, the other three variables also have a linear relationship with the grain output. Therefore, the multiple linear regression model was considered for fitting. Construct the regression model as follows:

$$Y = b_0 + b_1 X_1 + \cdots + b_7 X_7 + \varepsilon,$$

Table 2. Linear regression solution results

	Parameter estimation of regression	Standard Error	T-value	P value	Significance marker
constant	-2.589e+02	1.845e+02	-1.404	0.1738	
x_1	-9.288e-0.2	1.480e-01	-0.627	0.5366	
x_2	-1.613e+00	1.388e+00	-1.162	0.2570	
x_3	3.407e-01	6.136e-02	5.553	1.2e-05	***
x_4	1.807e-01	9.145e-02	1.976	0.0603	.
x_5	3.030e-03	1.353e-03	2.240	0.0351	*
x_7	4.470e+03	2.149e+03	2.080	0.0488	*
x_9	-7.391e+03	3.684e+03	-2.006	0.0567	.

From the above results, it can be seen that $p < 1.365e-14 < 0.01$ this equation is significant. In addition, the correlation coefficient R^2 is 0.9481, which is very high.

However, the significance of effective irrigation area and the amount of agricultural fertilizer applied in two variables basically did not exist. This indicates that the fitting effect is not good. Therefore, the method of stepwise regression is considered to solve this problem, and the results are as follows:

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

Table 3. Results of stepwise regression

	Parameter estimation of regression	Standard Error	T-value	P value	Significance marker
constant	-3.109e+02	1.778e+02	-1.749	0.0925	.
x_3	3.021e-01	5.135e-02	5.884	3.87e-06	***
x_4	1.159e-01	7.656e-02	1.513	0.1428	
x_5	2.645e-03	1.105e-03	2.393	0.0246	*
x_7	4.972e+03	2.062e+03	2.412	0.0235	*
x_9	-7.954e+03	3.606e+03	-2.206	0.0368	*

As can be seen from the above running results, $p < 2.839e-16 < 0.01$, indicating that this equation is significant. In addition, the correlation coefficient R^2 is 0.9486, although it is relatively high.It can be seen from this result that except for the number of combine harvesters, the remaining variables are all significant, and the fitting results are acceptable. The regression equation obtained is as follows:

$$Y = -310.9 + 0.301x_3 + 0.1159x_4 + 0.002645x_5 + 4972x_7 - 7954x_9$$

4.4. Results of Principal Component Analysis

The results of principal component analysis for each variable are shown in Table 4 below:

Table 4. Results of principal component analysis

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Principal component standard deviation	2.1271419	1.3473793	0.51805297	0.37849618	0.34016680	0.27680529	0.236354922
Variance contribution rate	0.6463904	0.2593473	0.03833984	0.02046562	0.01653049	0.01094588	0.007980521
Cumulative contribution rate of variance	0.6463904	0.9057376	0.94407748	0.96454310	0.98107360	0.99201948	1.000000000

From the calculation results, it can be seen that the cumulative contribution rate of the first two principal components has reached 90.50%, and the impact of the latter several principal components is small. The first principal component mainly reflects the effective irrigation area, while the second principal component reflects the amount of agricultural fertilizer applied. Therefore, two principal components are selected, and the other five principal components can be omitted to achieve the purpose of dimensionality reduction. In order to better determine the number of principal components, the gravel diagram is drawn as shown in Figure 5:

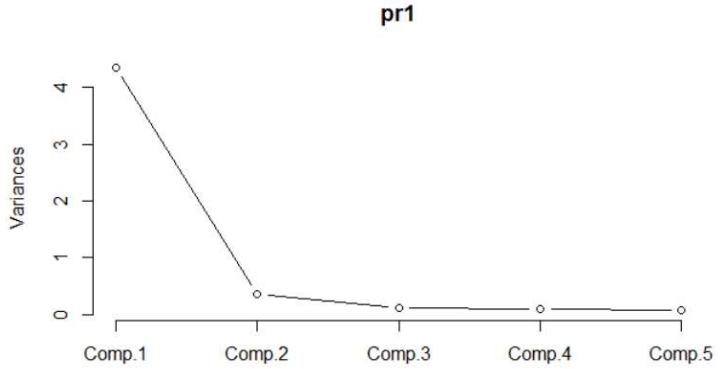

Figure 5. Gravel diagram

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

It can be seen from the rubble diagram and analysis results that it is more appropriate to select two principal component variables. Further observe the figure of principal component coefficients, as shown in Table 5 below:

Table 5. Table of principal component coefficients

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
x_1	0.448		0.233	0.414	-.277	0.320	0.622
x_2	0.416	0.227	-0.510	-0.305	-0.446	0.472	
x_3	0.447		-0.122	-0.137	0.698		-0.519
x_4	0.430	0.233	-0.175		-0.164	-0.816	0.190
x_5	0.426	-0.116	0.677		-0.430		-0.398
x_7	0.186	-0.654		-0.631	0.112		0.338
x_9	0.157	-0.661	-0.419	0.562	-0.113		-0.185

We can see from the coefficient of principal component: the first principal component mainly reflects the first 5 indicators (effective irrigation area, agricultural fertilizer application amount, total sown area of crops, number of combine harvesters, number of large and medium-sized agricultural tractors), we call the first principal component as resource input factor; The second principal component mainly reflects the sixth and seventh indicators (crop disaster rate, crop disaster rate), and we call the second principal component the disaster factor.

The comprehensive score is calculated according to the weighting method:

$$C = \frac{0.646390\,C_1 + 0.2593473\,C_2}{0.646390 + 0.2593473} = 0.71366188\,C_1 + 0.28633823\,C_2$$

The principal component scores and rankings of each region are as follows:

Table 6. Principal component ranking

Regions	Rankings	Regions	Rankings	Regions	Rankings
Beijing	22	Zhejiang	15	Hainan	19
Tianjin	20	Anhui	6	Chongqing	17
Hebei	14	Fujian	16	Sichuan	5
Shanxi	29	Jiangxi	8	Guizhou	12
Inner Mongolia	30	Shandong	3	Yunnan	11
Liaoning	27	Henan	2	Xizang	18
Jilin	23	Hubei	24	Shaanxi	26
Heilongjiang	1	Hunan	4	Gansu	28
Shanghai	21	Guangdong	9	Qinghai	31
Jiangsu	7	Guangxi	10	Ningxia	25
				Xinjiang	13

From the results of principal component analysis, we can get that Heilongjiang, Henan and Shandong ranked 1st, 2nd and 3rd respectively; Qinghai Province, Inner Mongolia Autonomous Region and Shanxi Province ranked 1st, 2nd and 3rd from the bottom respectively.

According to the results of principal component analysis, Heilongjiang, Henan and Shandong provinces are far ahead in grain production, while Qinghai, Inner Mongolia Autonomous Region and Shanxi provinces are low in grain production, which is mainly reflected in the effective irrigated area, the total sown area of crops, and the number of large and medium-sized tractors

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

for agricultural use. The number of large and medium-sized agricultural tractors, the effective irrigated area and the total sown area of crops in the top provinces such as Heilongjiang are much larger than those in the bottom regions such as Qinghai province.

Combined with the actual data, we found that the results of the principal component ranking are roughly the same as the rankings of grain output in 31 provinces, municipalities and autonomous regions in 2021. Since grain output is directly related to grain production level, the reliability of the results of the principal component analysis is high, and these factors determine the regional grain production level to a certain extent. However, there are also differences between the principal component ranking and the grain production ranking of some provinces and cities, which also shows that the grain production level is not only reflected in the grain production level. Because other conditions are equal, the larger the production area, the output will naturally increase. The degree of agricultural mechanization and the amount of fertilizer applied can also reflect the level of agricultural production in a place, so the grain yield is not the only factor to measure the level of food production.

4.5. Grey Correlation Analysis

The indexes selected in this paper are cost indexes except disaster rate and disaster rate. Therefore, the following analysis is based on the effective irrigation area, the conversion amount of agricultural fertilizer application, the total sown area of crops, the number of combine harvesters, the number of large and medium-sized agricultural tractors these five indicators are analyzed.

The standard 0-1 changes of the cost-type indicators are treated as follows:

$$b_{ij} = \frac{a_j^{\max} - a_{ij}}{a_j^{\max} - a_j^{\min}}$$

The correlation degree values of each region are obtained in the following table:

Table 7. Correlation degree table

Regions	Correlation value	Correlation ranking	Regions	Correlation value	Correlation ranking
Beijing	0.5957	18	Hubei	0.7828	27
Tianjin	0.6088	17	Hunan	0.5507	4
Hebei	0.8411	22	Guangdong	0.8364	12
Shanxi	0.9607	30	Guangxi	0.5504	11
Inner Mongolia	0.7810	29	Hainan	0.9472	20
Liaoning	0.6065	28	Chongqing	0.4404	14
Jilin	0.4745	23	Sichuan	0.7139	6
Heilongjiang	0.9877	2	Guizhou	0.4033	13
Shanghai	0.5145	21	Yunnan	0.5517	9
Jiangsu	0.7654	7	Xizang	0.9137	19
Zhejiang	0.6091	15	Shaanxi	0.7005	25
Anhui	0.6050	5	Gansu	0.9759	26
Fujian	0.9012	16	Qinghai	0.6964	31
Jiangxi	0.8505	10	Ningxia	0.6640	24
Shandong	0.7482	3	Xinjiang	0.6360	8
Henan	0.9971	1			

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

It can be seen from the above table that the correlation values of Shandong, Heilongjiang and Henan are in the top three respectively, indicating that the grain yield of these three regions is higher than that of other regions; While Qinghai, Shanxi and Inner Mongolia are at the bottom three, indicating that the grain yield of these three regions is lower than that of other regions.

5. Conclusion and Suggestions

5.1. Conclusion

Through empirical analysis, this paper draws the following conclusions:

(1) Firstly, cluster analysis is used to divide 31 regions. Shandong, Hunan, Heilongjiang and Henan can be divided into the first category, Inner Mongolia and Qinghai into the second category, and other regions into the third category.

(2) The linear regression model was constructed by using the knowledge of multiple regression to analyze and forecast the grain output in each province. The regression model was constructed by using the stepwise regression method. In the stepwise regression method, the total sown area of crops, the number of combine harvesters, the number of large and medium-sized agricultural tractors, the disaster rate and the disaster rate were selected as the independent variables to construct the regression model:

$$Y = -310.9 + 0.301x_3 + 0.1159x_4 + 0.002645x_5 + 4972x_7 - 7954x_9$$

(3) The principal component analysis method and grey correlation analysis method were used to analyze the grain output of 31 provinces in China and make a ranking. The results obtained by the two analysis methods were roughly the same, and the results were Heilongjiang, Shandong and Henan ranked the top three, indicating that their grain output was higher than that of other regions; While Qinghai, Inner Mongolia and Shanxi ranked the last three, indicating that their grain output was lower than that of other regions. It can be seen that the conclusions obtained by principal component analysis and grey correlation analysis are consistent with those obtained by cluster analysis above.

5.2. Policy Recommendations

Based on the above analysis, the following suggestions are given:

First, it is recommended to strictly control the occupation of arable land and the development and utilization of water resources, promote resource protection and efficient use of new technologies, new products and new projects, and constantly improve the efficiency of land and water resources use. We should continue to strengthen the construction of agricultural ecological protection. From this research, we can find that there are serious regional differences in China's agricultural development, which is of course related to the terrain and landform in some areas, temperature and climate. However, choosing reasonable cultivated crops, developing more efficient irrigation techniques, and using more agricultural machinery and fertilizers can all improve grain yield to some extent.

Secondly, it is suggested to devote ourselves to the development of efficient and water-saving irrigation methods such as sprinkler irrigation and drip irrigation in other dry areas where water is scarce. This requires not only the policy help of the state, but also a certain level of technology. The current situation that China's agricultural production is mainly concentrated in some plain areas makes it urgent to improve the level of mechanization. The degree of mechanization is particularly important in the context of the rapid development of science and technology in all countries in the world, and mechanization has greatly saved the time and manpower and material resources invested in agricultural production.

ISSN: 2692-7608 DOI: 10.6981/FEM.202312_4(12).0001

Third, not paying attention to the ecological environment has brought serious consequences to agriculture. For example, Qinghai and other "hard-hit areas" are prone to natural disasters. In addition to the climate and other factors, a large part of the reason is that the land is seriously desertified, and there is less and less land available for agricultural production. Overexploitation of agricultural resources, excessive use of agricultural inputs, overexploitation of groundwater, and overlapping of internal and external sources of agricultural pollution have brought about a series of increasingly prominent problems, and the sustainable development of agriculture faces major challenges.

References

[1] Wu Liping, Lin Jiaxiang. Analysis of agricultural economic development level in major grain-producing areas integrating principal components and clustering [J]. Journal of Yibin University, 2017(12):120-124.

[2] Shi H J. Influencing factors and cluster analysis of grain yield level [J]. Journal of Southwest Forestry University,2011(10) : 21-24.

[3] Peng W F. Grey analysis of influencing factors of grain production and grain yield prediction in Sichuan Province.]Journal of Sichuan Normal University, 2005(03):350-353.

[4] Wang Chunhui, Zhou Shenglu, Wu Shaohua et al. Prediction of grain yield in Jiangsu Province based on multiple linear regression model and grey correlation analysis [J]. Journal of Nanjing University. 2014(04):105-109.

[5] Lin Shan. Analysis on Influencing factors of grain output in Shaanxi Province [J]. Agricultural Staff, 2019(19):8.

[6] Liu Pengling, Wu Wenjun, Wan Yingying, Wan Guanghui. Analysis of influencing factors and grey prediction of grain yield: Based on data of main producing areas in Anhui Province [J]. Journal of Xian University of Architecture and Technology (Social Science Edition),2019,38(04):58-63.

[7] Yang Fanyu, Liu Liming, Yuan Chengcheng. Analysis and prediction of influencing factors of grain yield fluctuation in Hunan Province [J]. Chinese Agricultural Science Bulletin,20,36(29):153-160.

[8] XU Y, HUANG K M. Influencing factors of grain yield in Shandong Province based on Principal component analysis [J]. Journal of Qingdao Agricultural University (Natural Science Edition), 2021, 38(2):153-156.