Analysis of influencing factors of grain yield based on multiple linear regression

Victor Chang[1] and Qianwen Xu[2]

1. School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK

2. International Business School Suzhou, Xi'an Jiaotong-Liverpool University, Suzhou, China

Email: victorchang.research@gmail.com; iamarielxu@163.com

Prof. Victor Chang is a Full Professor of Data Science and Information Systems, School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK, since September 2019. Previously he was a Senior Associate Professor, Director of Ph.D. (June 2016–May 2018), Director of MRes (Sep 2017- Feb 2019) and Interim Director of BSc IMIS Programs (Aug 2018- Feb 2019) at International Business School Suzhou (IBSS), Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China. He served at XJTLU between June 2016 and August 2019. He was also a very active and contributing key member at Research Institute of Big Data Analytics (RIBDA), XJTLU and a key committee member at Research Center of Artificial Intelligence (RCAI), XJTLU. He was an Honorary Associate Professor at the University of Liverpool. He is still a Visiting Researcher at the University of Southampton, UK. Previously he worked as a Senior Lecturer at Leeds Beckett University, UK, for 3.5 years. Within 4 years, he completed Ph.D. (CS, Southampton) and PGCert (Higher Education, Fellow, Greenwich) while working for several projects at the same time. Before becoming an academic, he has achieved 97% on average in 27 IT certifications. He won a European Award on Cloud Migration in 2011, IEEE Outstanding Service Award in 2015, best papers in 2012, 2015 and 2018, the 2016 European award: Best Project in Research, 2016-2018 SEID Excellent Scholar, Suzhou, China, Outstanding Young Scientist award in 2017, 2017 special award on Data Science, 2017-2019 INSTICC Service Awards and numerous awards since 2012. He is a visiting scholar/Ph.D. examiner at several universities, an Editor-in-Chief of IJOCI & OJBD journals, Editor of FGCS, Associate Editor of TII, founding chair of two international workshops and founding Conference Chair of IoTBDS http://www.iotbd.org and COMPLEXIS http://www.complexis.org since Year 2016. He was involved in different projects worth more than £13 million in Europe and Asia. He has published 3 books as sole authors and the editor of 2 books on Cloud Computing and related technologies. He gave 18 keynotes at international conferences. He is widely regarded as one of the most active and influential young scientists and experts in IoT/Data Science/Cloud/security/AI/IS, as he has experience to develop 10 different services for multiple disciplines. He is the lead steering committee for IoTBDS, COMPLEXIS and FEMIB http://femib.scitevents.org/ to build up and foster active research communities globally.

Miss Qianwen Xu is an MSc in Business Analytics student from International Business School Suzhou, Xi'an Jiaotong-Liverpool University (XJTLU), Suzhou, China. She is expecting to graduate with distinction grade from both XJTLU and University of Liverpool master degree. She works under Prof Chang's under supervision. She is a hardworking, dedicated and resourceful student who can make things happen.

Abstract

Food security is a strategic issue affecting economic development and social stability and agriculture has always been at the forefront of national economic development. As a large agricultural country and a country with a large population, the production of grain is of great importance to China. Therefore, in order to ensure national food security and assist the food administrative department in making scientific and effective decisions, it is significant to study the law of variance in grain production and make accurate forecasting of its development trend. This paper constructs the stepwise regression model and principal component regression to analyze the influencing factors of grain yield respectively and compares these two models in terms of their accuracy in prediction. After conducting the two regressions, this paper concludes that the two models both explain the variance in grain yield ideally, but from the aspect of accuracy in prediction, the principal component regression is more effective than stepwise linear regression.

Keywords: grain yield, influencing factors, prediction, stepwise regression model, principal component regression

1 Introduction

As the country with the largest population, the Chinese national economy depends on the development of agriculture to a large extent. The basis of agriculture is grain output, therefore the grain production has become the eternal theme of the development and production of the nation and the study of grain production is of great significance to guide grain production. In order to provide a valuable reference model to forecast the grain yield in the future years, firstly, this paper chooses the datasets in which the period is from 1949 to 2017 and analyses the relationships between every single factor and the grain yield briefly. Next, this paper selects the datasets the period is from 1993 to 2012 to be the training sample and those period from 2013 to 2015 to be the testing sample and evaluate the relationship exists between these factors together and grain yield by using both stepwise regression model and principal component regression. Finally, the two models are compared with each other so that this paper can conclude the more valuable and effective one.

2 Literature Review

The literature on the factors that affect or may affect grain yield will be present in this part. And this paper will select the variables for conducting the empirical analysis by referencing the literature and data available.

There are a number of discussions about factors may affect grain yield. Doberman (1994) employs a factor analysis and multiple linear regression to analyze the contribution of five factors to changes in rice yield, including soil fertility, application of nitrogen fertilizer, application of phosphate fertilizer, reserved land resources and planting rate. Timsina et al (2001) use standard procedures of split-split plot design for rice and split-split-split design for wheat to discuss the effects of planting system selection between rice and wheat, sowing time, mechanization, and organic fertilizer on yield growth. Samapundo (2005) uses the Gompertz equation (Marin et al., 1996). to quantitatively analyze the impact of water resources and temperature on production growth. While Downing (1992), Wheeler and Von Braun (2013) consider climate change impacts crop productivity, Isik and Devadoss (2006) consider the

effect is limited by employing an econometric model to estimate stochastic production functions. By analyzing a large panel dataset of Indian agriculture with a random effects model, Kanwar (2006) finds the productivity of food crops is influenced by rainfall, specifically irrigation, fertilizer and improved seed.

3 Model Background

The two models this paper employs are both based on multivariable linear regression whose equation is as followed:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \varepsilon_i$$

where

i=1,2,3…;

y_i: dependent variable;

x_i: independent variable;

β_i: parameter.

In this part, this paper will introduce the stepwise regression model and principal component regression, and then explain the reason this paper chooses the two models.

3.1 Stepwise Regression Model

Stepwise regression is a procedure that predictor variables are investigated sequentially, variables are entered into or moved from the initial model one by one (Mundry and Charles,2009). In SPSS, the initial model contains no predictors and then the strongest variable is added into the model if the coefficient is significant. After that, the second strongest variable will enter into the model, however, some previously entered predictors will not be significant as more variables enter into the model, therefore SPSS will remove them from the model. (Stepwise Regression in SPSS, 2018). Compared to just conducting multivariable linear regression in SPSS by default, stepwise regression removes the factors that have little impact on dependent variables. This model helps to increase the accuracy in prediction as it reduces the degree of difficulty and increases the calculation efficiency and stability of the regression equation. In the case of analyzing the variance in grain yield, there are plenty of potential

factors that may affect the grain yield more or less. Including as many variables as possible in the model may enable the model to explain the proportion of variance in grain yield more specifically, but doing so will also complicate the model unnecessarily. Therefore, this paper employs the stepwise regression model as one of the analysis methods to avoid this problem.

3.2 Principal Components Analysis

Principal components analysis (PCA) is a statistical method and its purpose is to reduce the dimension of the dataset. PCA converts a set of correlated variables into a smaller number of uncorrelated variables and these uncorrelated variables are called principal components. The relationship between principal components and dependent variables is linear (Einasto, et al, 2011). Principal component analysis has applied in the technology of facial recognition to identify and verify facial features (Karamizadeh, et al.,2013). The key advantage of principal component analysis is to solve the multicollinearity problem which results from the correlation relationship among explanatory variables (Jolliffe, 2011). In addition, it is able to reduce the dimension and maintain most of the information at the same time. In the case of analyzing the variance in grain yield, it is much possible that explanatory variables are mutually correlated. For example, grain yield is expected to partly depend on the area of cultivated crops, the amount of fertilizer used and the use of pesticides. In the meantime, the amount of fertilizer used and the use of pesticides are each expected to have a positive correlation with the area of cultivated crops based upon economic assumptions. Hence, the principal component analysis is selected by this paper and compared with stepwise linear regression.

4 Analysis of the current situation

4.1 The trend of China's total grain output

According to Figure 1, it shows more specifically that the general trend of grain yield from 1949, the year of new China established, to 2017 is going up, but there were still several drops during 69 years. Among these drops, two of them were of importance. One was the decrease from 1959 to 1961, the three years was the famous economic hardship of China, resulting from the

activity called Great Leap Forward violating the law of development and undermining agricultural production (Liu, 2010). The other occurred from 1998 to 2003. The grain yield of these five years dropped by nearly about 16%. The main reasons were the drought and the rapid reduction in planting area in order to respond to the policy of returning farmland to forests and grass (Chen and Li, 2013).

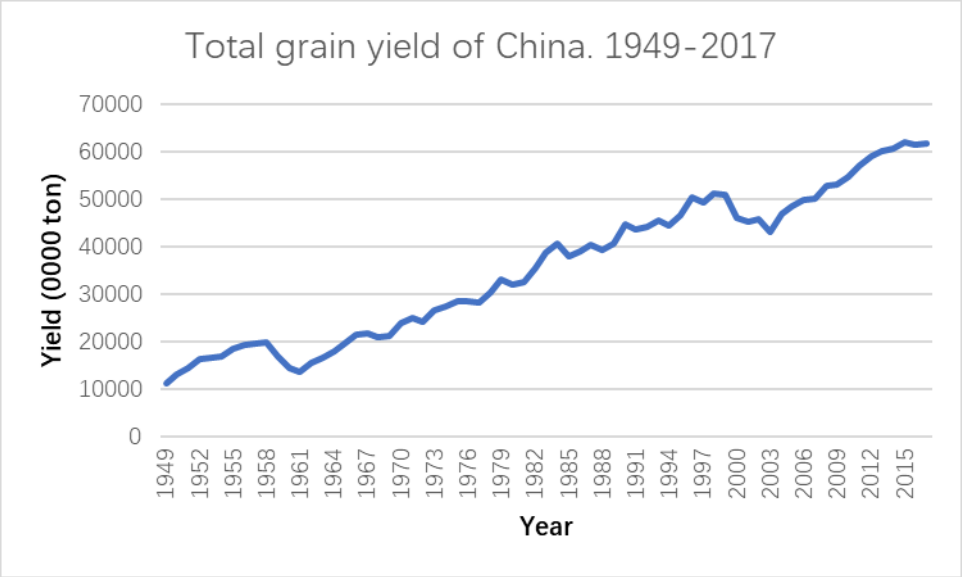

Figure 1 Total grain yield of China. 1949-2017

However, as the country attached importance to agriculture and increased investment in agricultural water conservation and machinery, total grain output has kept increasing in recent years. In the following paragraphs, this paper will discuss the impacts of single factors on the variance in grain yield.

4.2 Analysis of the influence of single factors

The primary factor this paper will discuss is the sown area as other factors are controlled, the grain yield will increase as the sown area increases. However, Figure 2 shows that the trend of the sown area dropped a little while the grain yield kept growing. It indicates that grain yield is no longer purely depends on the sown area. With the development of technology and acknowledge of grain, the sown area can be utilized in a much more efficient way.

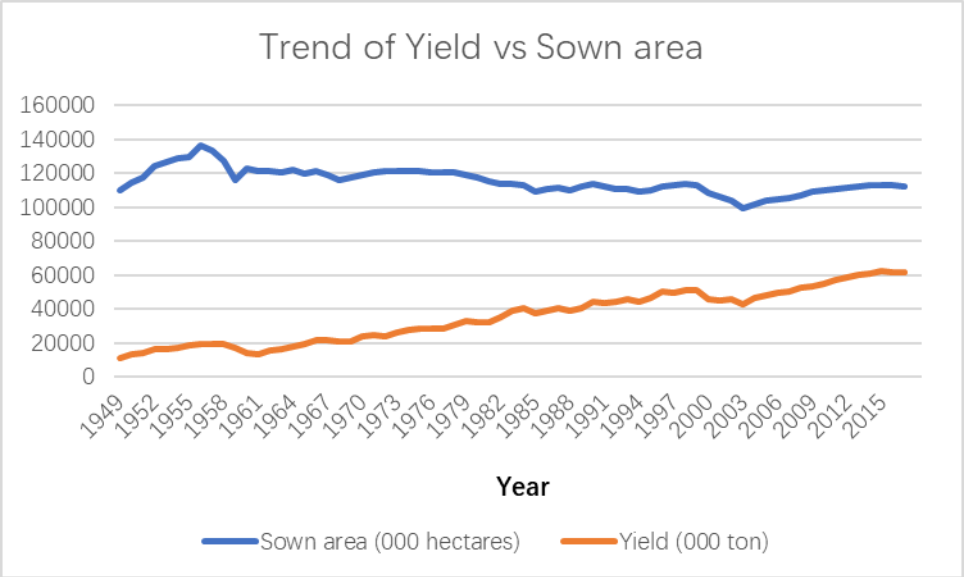

Figure 2. Trend of Yield vs Sown area. 1949-2017.

According to Figure 3 and Figure 4, these trends show that the relationships between grain yield and effective irrigation area or waterlog control area are both positive, which are in line with expectations. With the rapid development of technology in irrigation and flood prediction and control, more and more farmlands can be irrigated more effectively and protected from waterlogging.

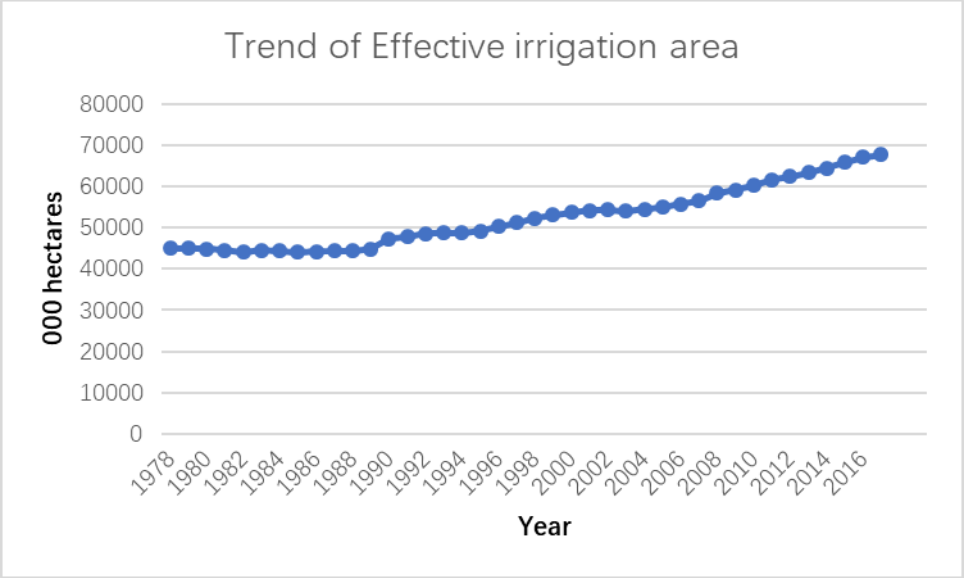

Figure 3. Trend of Effective irrigation area.1978-2016

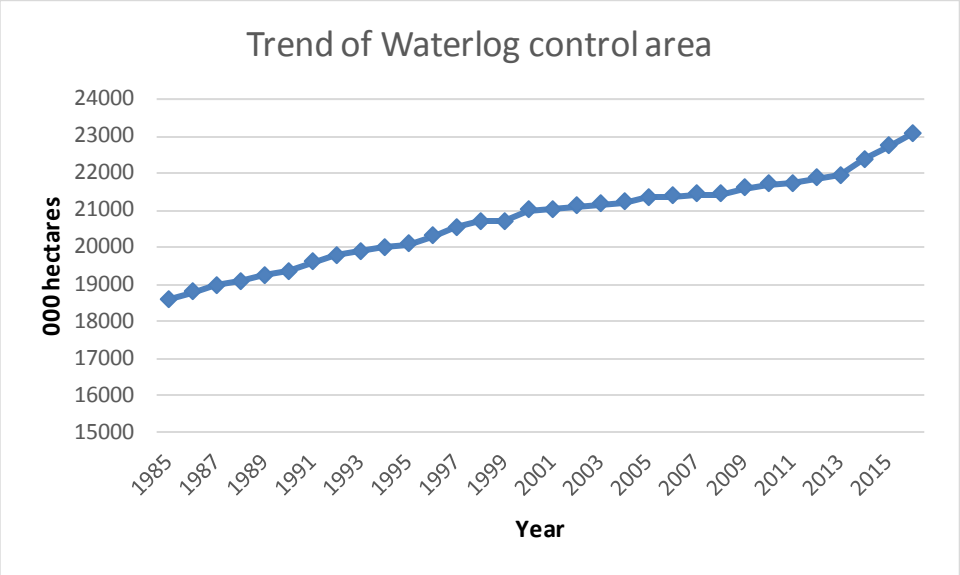

Figure 4. Trend of Waterlog control area. 1985-2017.

Figure 5 shows that the areas affected by disasters fluctuated greatly during the whole time. This is on account of the frequency of disaster occurrence varies from year to year, but the overall trend is negative and matches the realistic meaning.

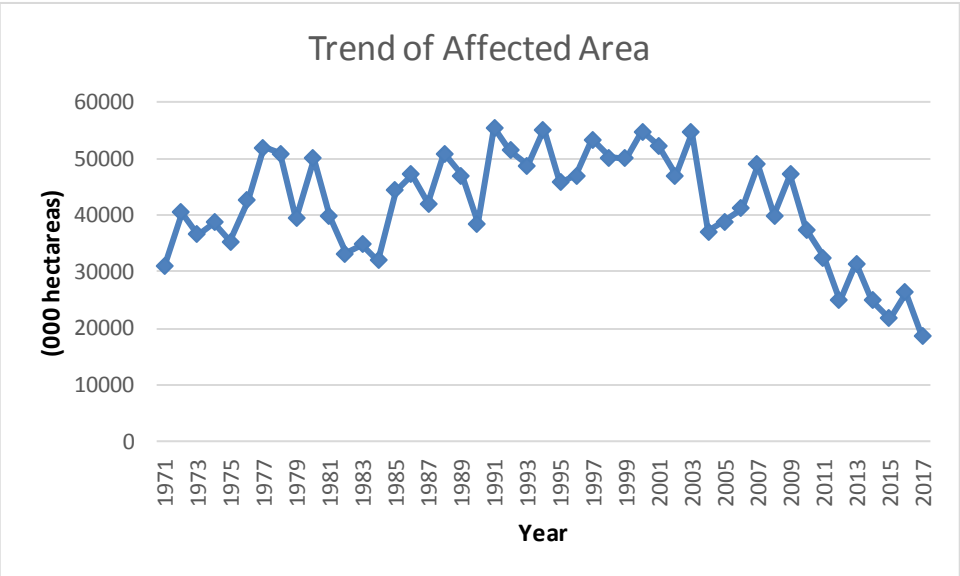

Figure 5. Trend of Affected Area. 1971-2017

Aside from the above factors, the usage of fertilizer can be an important factor that will impact the grain yield. In the food industry, it is of great importance to make a scientific fertilization scheme because it not only can improve the fertilization effect but also reduce fertilization

cost (Lin and Su, 2004). There are different kinds of chemical fertilizers. As shown in Figure 6, the total usage has kept going up from 1983 to 2016. The use of compound fertilizer has increased while other fertilizers remained constant. This means that the knowledge of scientific fertilization has been popularized.

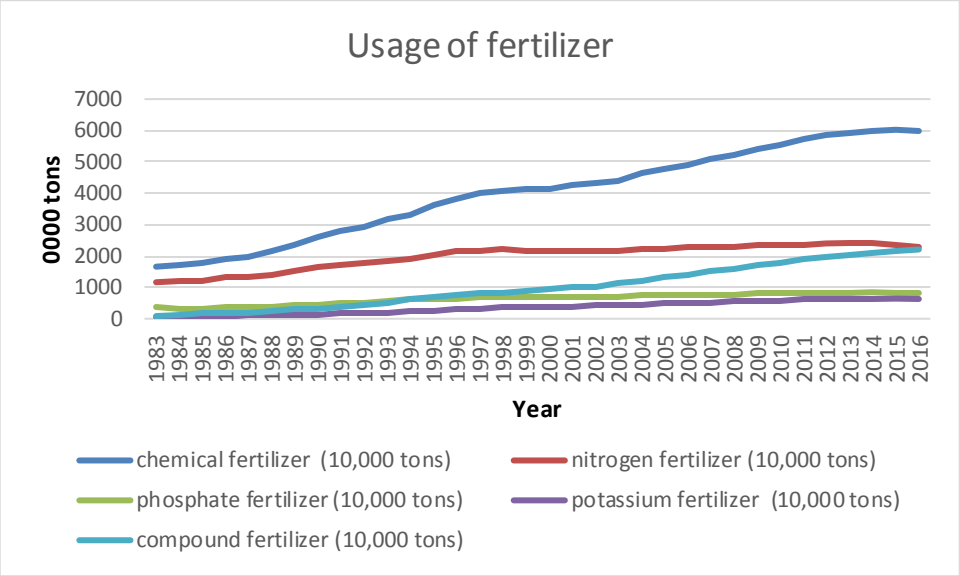

Figure 6. Usage of fertilizer. 1983-2016.

From Figure 7 to Figure 10, it can be seen that the trend of power of agricultural machinery, diesel consumption, pesticide usage and agricultural plastic film usage all kept increasing as the grain yield kept increasing in the same period. Therefore, this paper assumes that the relationship between each of these four factors and grain yield is positive.

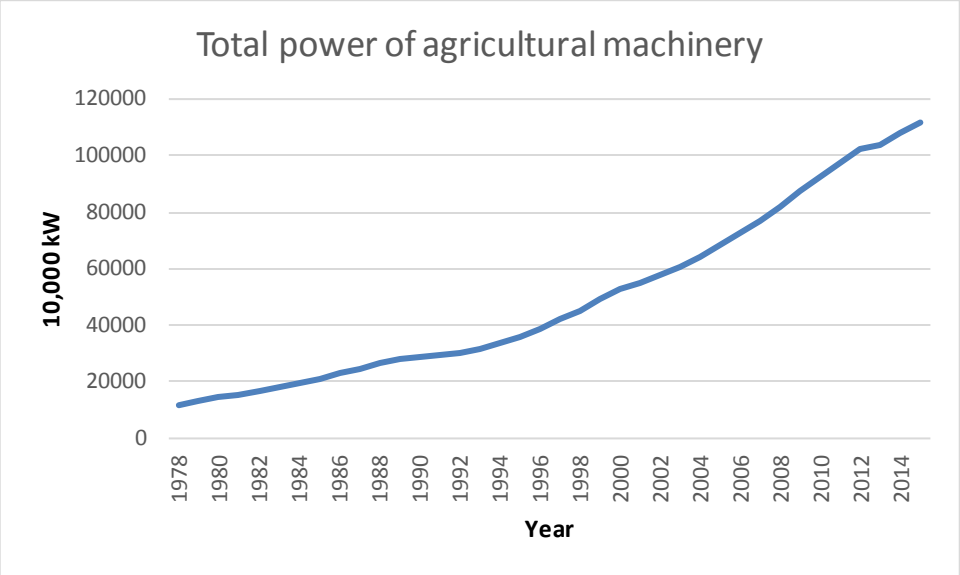

Figure 7. Total power of agricultural machinery. 1978-2014.

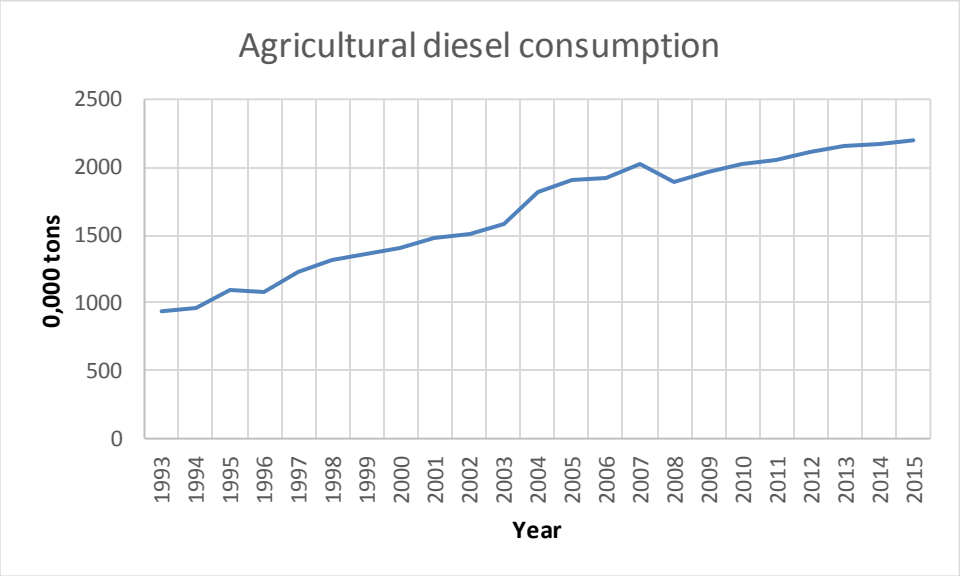

Figure 8. Agricultural diesel consumption. 1993-2015.

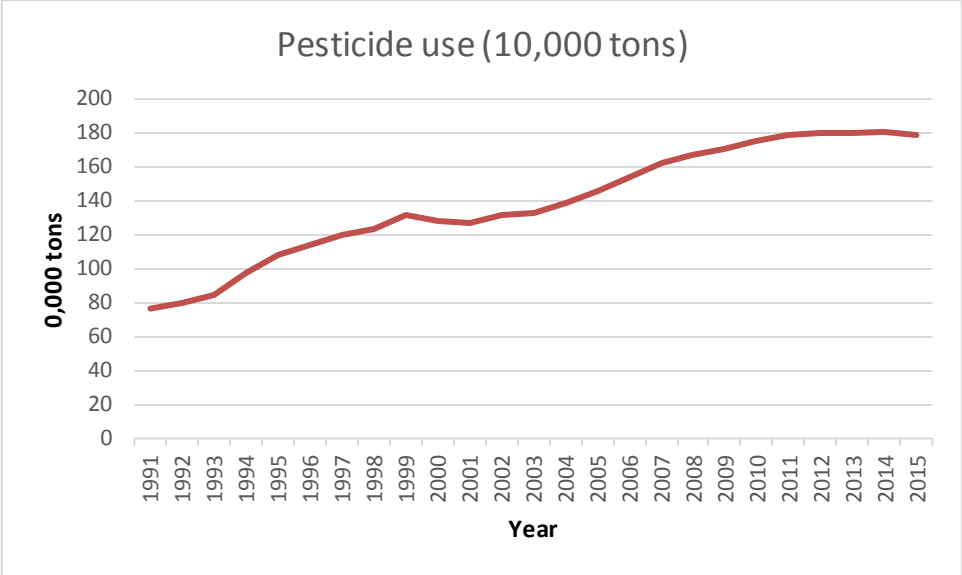

Figure 9. Pesticide use. 1991-2015.

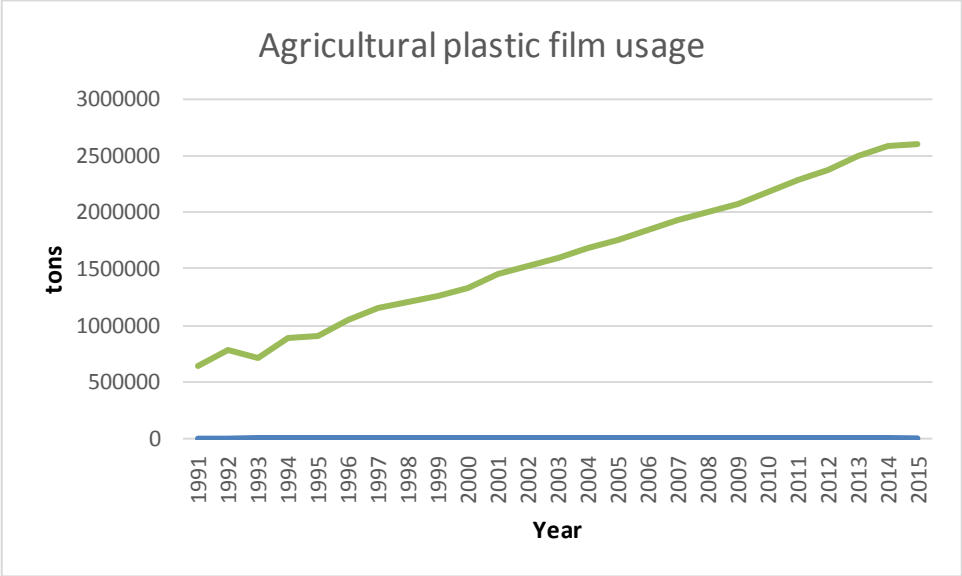

Figure 10. Agricultural plastic film usage. 1991-2015.

The above paragraphs discuss the current situation of Chinese agriculture development in general as well as the influence of single factors. In the following parts, this paper will conduct stepwise regression and principal component regression respectively to analyze the influence of factors selected based on the literature and the analysis in this part.

5 Data Management

5.1 Data source

The dataset this paper uses is from the website National Bureau of Statistics of China (NBS). As an agency directly under the State Council, the NBS (data.stats.gov.cn) is in charge of statistics and economic accounting in China. This paper chooses the dataset from NBS because the dataset from NBS is much convenient to obtain and accountable. The period of the original dataset obtained from NBS is from 1949 to 2017. In consideration that the data of some variables were absent until the 1990s, this paper chooses the dataset from 1993 to 2015 to conduct the analysis. And among them, this paper selects the datasets the period is from 1993 to 2012 to be the training sample and those period from 2013 to 2015 to be the testing sample.

5.2 Variables Selection

The evaluation should include the factors that are currently able to explain the change of China's total grain output so that the results of this analysis can provide practical, feasible and effective recommendations for improving China's grain output. Based on the reference to other literature and in consideration of the availability of data, this paper chooses nine factors as independent variables to study their relationship with grain yield, which is independent variables, and these variables are divided into two categories. One is the group of land factors and the other is the group of technology factors. The group of land factors consists of sown area (X_8), effective irrigation area (X_2), affected area (X_9) and waterlog control area (X_4). And the group of technology factors includes chemical fertilizer (X_3), pesticide use (X_6), total power of agricultural machinery (X_1), agricultural diesel consumption (X_5) and agricultural plastic film usage (X_7). These variables are described in detail as below:

Y-grain yield (0,000 tons): The amount of grain that has harvested in one calendar year. The unit of measurement is ten thousand tons.

X1- Total power of agricultural machinery (0,000 kw): the amount of power the agricultural machinery use in one calendar year and the energy provided for the machinery usually comes from electricity and diesel. The unit of measurement is kilowatt.

X2- Effective irrigation area (000 hectares): the area of cultivated land that has been irrigated effectively in one calendar year. The cultivated lands have a certain source of water and a complete irrigation system. The unit of measurement is a thousand hectares.

X3- chemical fertilizer (0,000 tons): the amount of fertilizer used in one calendar year. The types of chemical fertilizers include nitrogen fertilizer, phosphate fertilizer, potassium fertilizer as well as compound fertilizer. The unit of measurement is ten thousand tons.

X4- Waterlog control area (000 hectares): the area of cultivated land that the frequency of waterlogging's occurrence is once every three years or above. The cultivated lands are protected by waterlogging facilities. The unit of measurement is a thousand hectares.

X5- Agricultural diesel consumption (0,000 tons): the amount of diesel spent on agricultural activities in one calendar year. The unit of measurement is ten thousand tons.

X6- Pesticide use (0,000 tons): the amount of pesticide spent on agricultural activities in one calendar year. The unit of measurement is ten thousand tons.

X7- Agricultural plastic film usage (tons): the amount of plastic film spent on agricultural activities in one calendar year. The unit of measurement is tons.

X8- Sown area (000 hectares): the area under crops. The unit of measurement is hectare.

X9- Affected Area (000 hectares): the area of cultivated lands that have affected by natural disasters in one calendar year. The types of disasters consist of drought, flood, snow and hail disaster. The unit of measurement is a thousand hectares.

6 Empirical Analysis

After analyzing the influence of single factors on grain yield, this paper will construct the stepwise regression model and principal component regression to analyze the influencing factors of grain yield respectively in the following paragraphs.

6.1 Stepwise linear regression

First of all, this paper conducts the stepwise linear regression model by importing the dataset into SPSS Statistics to analyze the relationship between the selected explanatory variables and grain yield. Stepwise linear regression is a method of regressing multiple variables while

simultaneously removing those that are not significant to the model. The previously entered predictors will be tested again and will be removed from the model if there is enough evidence to show that they are not significant any more as more variables enter into the model. The result is as following:

Table 1

Variables Entered/ Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Effective irrigation area (000 hectares) X2	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051).
2	Sown area (000 hectares) X8	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051).
3	chemical fertilizer (0,000 tons) X3	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051).
4	.	Effective irrigation area (000 hectares) X2	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051).
5	Affected Area (000 hectares) X9	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .051).

a. Dependent Variable: grain yield (0,000 tons)

From Table 1, which shows the variables entered or removed, it can be seen that the criteria for variables to enter or remove is:

if $p<0.05$, enter;

if $p>0.05$, remove.

The p-value denotes the significance level for which the null hypothesis is rejected. If a p-

value is smaller than the size (e.g. 0.05), the null hypothesis will be rejected. In the case of table 1, the null hypothesis is the variable entered is not significant to the model. If a p-value is smaller than 0.05, then the null hypothesis will be rejected and the variable will be allowed to enter; If a p-value is greater than 0.05, then the null hypotheses will be accepted and the variable will not be allowed to enter or will be removed from the model.

The variable 'effective irrigation area' is the first one to enter the model but as variable 'Sown area' and 'chemical fertilizer' added into the model, the variable 'effective irrigation area' is proved to be insignificant and is removed from the model. Finally, among the nine independent variables, three variables are significant and selected to enter the model. The three variables are 'Sown area (X8)', 'chemical fertilizer (X3)' and 'Affected Area (X9)'.

Table 2

Model Summary

Model	R Square	Adjusted R Square	Std. Error of the Estimate	F	Sig.
1	0.588	0.565	2807.2323	25.7	.000[a]
2	0.915	0.905	1310.2	91.807	.000[b]
3	0.957	0.949	961.466	118.845	.000[c]
4	0.955	0.949	958.71844	178.837	.000[d]
5	0.983	0.98	605.431	307.841	.000[e]

a. Predictors: (Constant), Effective irrigation area (000 hectares) X2

b. Predictors: (Constant), Effective irrigation area (000 hectares) X2, Sown area (000 hectares) X8

c. Predictors: (Constant), Effective irrigation area (000 hectares) X2, Sown area (000 hectares) X8, chemical fertilizer (0,000 tons) X3

d. Predictors: (Constant), Sown area (000 hectares) X8, chemical fertilizer (0,000 tons) X3

e. Predictors: (Constant), Sown area (000 hectares) X8, chemical fertilizer (0,000 tons) X3, Affected Area (000 hectares) X9

SPSS constructs five models while regressing variables stepwise, and Table 2 shows that model 5 is the most efficient as its adjusted R^2 (0.98) is the largest. Moreover, the lowest

value of standard error (605.431) and the largest F-value (307.841) of model 5 also prove that it is the best model among these five models. And the possibility of significance of model 5 is less than 0.05, meaning that the null hypothesis 'all parameters of variable ''Sown area', 'chemical fertilizer' and 'Affected Area' equals 0' should be rejected, as there is not enough evidence to prove that all variables together are not significant to the dependent variable 'grain yield'. Model 5's R^2 (0.983) indicates that the variance of the three variables together can explain about 98% of the variance in the dependent variable 'grain yield'.

Table 3

Excluded Variables

Model		Beta In	t	Sig.
5	Total power of agricultural machinery (0,000 kW) X1	-0.257	-1.084	0.296
	Waterlog control area (000 hectares) X4	-0.334	-1.955	0.069
	Agricultural diesel consumption (0,000 tons) X5	-0.11	-0.558	0.585
	Pesticide use (0,000 tons) X6	0.082	0.223	0.827
	Agricultural plastic film usage (tons) X7	-0.514	-0.969	0.348
	Effective irrigation area (000 hectares) X2	-0.276	-1.603	0.13

SPSS also summarizes the outcomes of excluded variables and the results are shown in Table 3. As is shown in Table 3, the possibilities of the significance of all these six variables excluded from the model are greater than 0.05. It means the null hypothesis 'each of the six variable's estimated parameter equals 0' is accepted and there is not enough evidence to prove that these variables have a strong relationship with the independent variable 'grain yield', therefore the model excludes these six variables.

In summary, according to Table 4, there are unstandardized coefficients and standardized coefficients. As the measurement units of variables are not consistent, so the standardized coefficients are employed and the equation obtained from the stepwise regression model is:

$$y = 0.594x_8 + 0.686x_3 - 0.248x_9 \quad (1)$$

Table 4

Coefficients

Model		Unstandardized Coefficients	Std. Error	Standardized Coefficients	t	Sig.
		B		Beta		
5	(Constant)	-29846.89	4865.913		-6.134	0
	Sown area (000 hectares) X8	0.632	0.036	0.594	17.633	0
	chemical fertilizer (0,000 tons) X3	3.729	0.265	0.686	14.063	0
	Affected Area (000 hectares) X9	-0.13	0.025	-0.248	-5.16	0

6.2 principal component regression

After applying the stepwise regression model, this paper constructs the principal component regression model as well. The results are as followed:

Table 5.

Model Summary[b]

Model	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson	F	Sig.
6	0.989[a]	0.979	616.089	2.081	99.7	.000[b]

a. Predictors: (Constant), X_1-X_9;

b. Dependent Variable: grain yield (0,000 tons)

From Table 5, the value of R^2 (0.989) indicates that the fitness of the model is good and the nine variables together can explain the variance in the dependent variable 'grain yield' by 98.9%. In addition, the possibility of significance is close to 0, meaning that the null hypothesis 'the coefficients of all the nine variables are equal to 0' is rejected and all of the explanatory

variables together have a strong relationship with the dependent variable 'grain yield'. In the meantime, the value of Durbin-Watson (DW) is shown in the table and it is used to test whether there is an autocorrelation problem. An autocorrelation problem arises if different error terms are correlated. If the value of DW is close to 2, then there is no autocorrelation problem as in this case of grain yield.

Table 6

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics
		B	Std. Error	Beta			VIF
6	(Constant)	35265.957	40320.659		0.875	0.402	
	X1	-0.141	0.139	-0.73	-1.019	0.332	466.511
	X2	0.263	0.618	0.255	0.426	0.679	324.003
	X3	8.189	4.268	1.507	1.919	0.084	559.421
	X4	-4.095	3.155	-0.578	-1.298	0.224	180.181
	X5	0.771	3.589	0.072	0.215	0.834	101.387
	X6	-37.903	63.76	-0.246	-0.594	0.565	155.53
	X7	0.004	0.007	0.409	0.514	0.619	574.366
	X8	0.564	0.099	0.529	5.674	0	7.9
	X9	-0.107	0.037	-0.204	-2.904	0.016	4.461

When the nine explanatory variables are analyzed respectively, Table 6 shows that except variables 'Sown area (X8)' and 'Affected Area (X9)', other variables' possibility of significance are all greater than 0.05. It seems that their relationships respectively with the dependent variable 'yield field' are not significant but the value of VIF (greater than 10) indicates that the problem called multicollinearity exists. Multicollinearity is used to describe the situation when an exact or approximate linear relationship exists between explanatory variables and this problem will lead to unreliable regression estimates. The Variance Inflation Factor (VIF) can be used to detect multicollinearity and if the value of VIF is 10 or more, then the multicollinearity problem may exist. In order to solve this problem, this paper uses SPSS to analyze the principal components and the outcome is as followed:

Table 7

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.759
Bartlett's Test of Sphericity	df	36
	Sig.	0

The value of Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMS) is to measure whether the dataset is suitable to conduct the factor analysis and the critical value is 0.5. If KMS is greater than 0.5, then the dataset is suitable to do the analysis and if KMS is less than 0.05, then the dataset is not. According to table 7, the value of KMS is 0.759, meaning that the dataset this paper chose is suitable to conduct the factor analysis and the possibility of significance is around 0, indicating that the explanatory variables are related to each other.

Table 8

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.425	82.497	82.497	7.425	82.497	82.497
2	1.053	11.698	94.196	1.053	11.698	94.196
3	0.426	4.728	98.924			
4	0.049	0.541	99.465			
5	0.029	0.318	99.783			
6	0.014	0.151	99.933			
7	0.004	0.044	99.977			
8	0.001	0.013	99.99			
9	0.001	0.01	100			

As is shown in Table 8, the method of Principal Component Analysis extracts two principal components (F1&F2) and the percentage of their contribution to explain the dependent variable is up to about 94% cumulatively. According to Table 8 and Table 9, the equation used to describe the relationship between the principal components and the original variables can

be figured out:

$F1 = 0.36X_1 + 0.36X_2 + 0.36X_3 + 0.36X_4 + 0.36X_5 + 0.36X_6 + 0.37X_7 - 0.09X_8 - 0.28X_9$ (2)

$F2 = 0.05X_1 + 0.11X_2 + 0.06X_3 - 0.13X_4 - 0.13X_5 + 0.08X_6 - 0.03X_7 + 0.93X_8 - 0.29X_9$ (3)

where

the coefficient=the coefficient of component/ (Initial Eigenvalues)$^{1/2}$,

the coefficient of component is from Table 9;

Initial Eigenvalues are from Table 8.

Table 9

Component Matrix[a]

Variables	Component	
	1	2
Total power of agricultural machinery (0,000 kW) X1	0.992	0.054
Effective irrigation area (000 hectares) X2	0.98	0.111
chemical fertilizer (0,000 tons) X3	0.994	0.061
Waterlog control area (000 hectares) X4	0.975	-0.131
Agricultural diesel consumption (0,000 tons) X5	0.979	-0.132
Pesticide use (0,000 tons) X6	0.982	0.078
Agricultural plastic film usage (tons) X7	0.997	-0.028
Sown area (000 hectares) X8	-0.25	0.951
Affected Area (000 hectares) X9	-0.751	-0.297

After extracting out the two principal components, the relationship between the principal components and dependent variables should be regressed again. However, during the process of extracting principal components, the explanatory variables have been standardized as the dependent variable has not. Therefore, before conducting the second multivariable linear regression, the dependent variable 'grain yield' needs to be standardized first.

Table 10

Model Summary

Model	R Square	Adjusted R Square	F	Sig.
6	0.989	0.979	99.7	.000[b]
7	0.976	0.973	348.378	.000[b]

a. Predictors: (Constant), X_1-X_9;

b. Predictors: (Constant), F2, F1

After conducting the second linear regression (model 7), this paper compares it with model 6 and the result is shown in Table 10. The R2 and adjusted R2 of two models are similar but the F-value of model 7 (348.378) is about 3.5 times bigger than the value of model 6 (99.7), meaning that model 7 is more accurate than model 6.

Table 11

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	VIF
		B	Std. Error	Beta			
1	(Constant)	4.89E-15	0.036		0	1	
	F1	0.263	0.014	0.716	19.124	0	1
	F2	0.664	0.036	0.681	18.194	0	1

According to Table 11, F1 and F2's possibilities of significance are all close to 0, meaning that they have a significant relationship with the dependent variable 'grain yield'. When VIF of the two factors are all equal to 1, it mean that there is no multicollinearity problem. The equation on the relationship between the standardized dependent variable and principal components is:

$$Y = 0.263F1 + 0.664F2 \quad (4)$$

Then this paper plugs equation (2)(3) into equation (4), then equation (5) which represents the relations between the standardized dependent variable and standardized explanatory variables is as followed:

$$Y = 0.13X_1 + 0.16X_2 + 0.13X_3 + 0.01X_4 + 0.01X_5 + 0.14X_6 + 0.08X_7 + 0.59X_8 - 0.26X_9 \quad (5)$$

In order to get the equation on the relationship between the dependent variables and original explanatory variables, the standardized equation should be converted into the unstandardized equation:

$$y = -31679.37 + 0.025x_1 + 0.171x_2 + 0.729x_3 + 0.062x_4 + 0.091x_5 + 22.155x_6 + 0.001x_7 + 0.626x_8 - 0.137x_9 \quad (6)$$

where

the transfer equation is: $\frac{y-\bar{y}}{S_y} = 0.13\frac{x_1-\overline{x_1}}{s_{x_1}} + 0.16\frac{x_2-\overline{x_2}}{s_{x_2}} + 0.13\frac{x_3-\overline{x_3}}{s_{x_3}} + 0.01\frac{x_4-\overline{x_4}}{s_{x_4}} + 0.01\frac{x_5-\overline{x_5}}{s_{x_5}} + 0.14\frac{x_6-\overline{x_6}}{s_{x_6}} + 0.08\frac{x_7-\overline{x_7}}{s_{x_7}} + 0.59\frac{x_8-\overline{x_8}}{s_{x_8}} - 0.26\frac{x_9-\overline{x_9}}{s_{x_9}}$;

the value of $\bar{y}$, S_y, s_{x_i}, x_i is from Table 12.

Table 12

Descriptive Statistics

Variables	Mean	Std. Deviation
Grain yield (0,000 tons) Y	49550.453	4257.36827
Total power of agricultural machinery (0,000 kW) X1	62343.1835	22034.1347
Effective irrigation area (000 hectares) X2	54715.4445	4114.24589
chemical fertilizer (0,000 tons) X3	4520.891	783.34687
Waterlog control area (000 hectares) X4	20997.6	601.267
Agricultural diesel consumption (0,000 tons) X5	1582.07	396.57946
Pesticide use (0,000 tons) X6	138.646	27.64535
Agricultural plastic film usage (tons) X7	1563133.297	493620.3491
Sown area (000 hectares) X8	108213.877	3999.48294
Affected Area (000 hectares) X9	45333.05	8136.834

6.3 Model Evaluation

After conducting the stepwise regression model and principal component regression, this paper will compare the two models by the accuracy in prediction with the sample from 2013 to 2015. And the equation about the percentage of accuracy is as followed:

Accuracy%=1-|predicted value – actual value|/actual value

In order to get the predicted value, this paper puts the value of explanatory variables from 2013 to 2015 into equation (1) and equation (6) respectively, and the results of the two models are in Table 16. From table 16, this paper can conclude that in predicting the grain yield, principal component regression is more effective than stepwise linear regression from the aspect of accuracy.

Table 13

Year / grain yield	2013	2014	2015
Actual Value (0,000 tons)	60193.84	60702.61	62143.92
Stepwise (0,000 tons)	62782.34	64897.46	66058.24
Principal Component (0,000 tons)	59010.67	60822.80	61964.42
Stepwise (Accuracy%)	95.70%	93.09%	93.70%
Principal Component (Accuracy%)	98.03%	99.80%	99.71%

7 Results Discussion

In practice, there may be some limitations with principal component regression, comparing to the stepwise regression model. Firstly, principal component regression is more complex to conduct and the coefficients of the final unnormalized equation will not be given directly by SPSS or other tools. Further calculation should be carried out in principal component regression while the final regression result is easy to get from the stepwise regression model. Secondly, also because of the complex process, stepwise regression provides easier-to-interpret coefficients and might be easier to explain to others than PCA. However, with a multicollinearity problem in the regression model like this case, the regression estimates may be unreliable and biased. This disadvantage is especially prominent when there are many variables and the method of PCA is a good method to deal with it. Therefore, this paper employs the model of principal component regression as there is no multicollinearity problem

and the value of accuracy is greater than the stepwise regression model's.

$$y = -31679.37 + 0.025x_1 + 0.171x_2 + 0.729x_3 + 0.062x_4 + 0.091x_5 + 22.155x_6 + 0.001x_7 + 0.626x_8 - 0.137x_9 \quad (6)$$

From the coefficients of explanatory variables in equation (6), it indicates that except the negative influence of 'affected area' on grain yield, the other eight variables have a positive impact on grain yield, and the detailed description is as followed:

Pesticide use (0,000 tons): If other variables are controlled, 10,000 extra tons of pesticide is used, the output of grain will increase by 221,550 tons.

Chemical fertilizer (0,000 tons): If 10,000 more tons of chemical fertilizer is applied, the grain yield will go up by 7,290 tons.

Sown area (000 hectares): If the sown area expands by 1,000 hectares, the grain yield will increase by 6,260 tons.

Effective irrigation area (000 hectares): If the area of farmland which is irrigated effectively increases by 1,000 hectares, the grain yield will go up by 1,710 tons.

Agricultural diesel consumption (0,000 tons): The grain yield will increase by 910 tons as the agricultural diesel consumes extra 10,000 tons.

Waterlog control area (000 hectares): If extra 1,000 areas of farmlands' waterlog are controlled effectively, the output of grain will increase 620 tons.

Total power of agricultural machinery (0,000 kW): If 10,000 extra kilowatts of power of agricultural machinery is used, the grain yield will go up by 250 tons.

Agricultural plastic film usage (tons): If one extra ton of agricultural plastic film is used, the grain production will increase by 10 tons.

Affected Area (000 hectares): The area of farmland affected has negative impacts on grain yield and 1,370 tons of grain yield will decrease as 1,000 hectares of farmland affected.

8 Conclusion

This paper firstly discusses the trend of total grain yield of China from 1949 to 2017 as well as the development of several single factors that may have an influence on grain yield. Then

based on multivariable linear regression, this paper constructs the stepwise regression model and principal component regression to analyze the influencing factors of grain yield respectively and compares these two models in terms of their accuracy in prediction. The datasets this paper chooses are from 1993 to 2015, and among them, the dataset from 1993 to 2012 to be the training sample and that from 2013 to 2015 to be the testing sample. After conducting the analysis, from the aspect of goodness of fit, the stepwise model explains the variance of grain yield better than the principal component regression but from the perspective of prediction, the principal component regression is more effective than stepwise linear regression.

There are several limitations with this research. Firstly, due to the lack of relevant data until the 1990s, this paper only studies the changes in grain production since 1990, and does not consider the longer-term span. This may affect the research depth to some extent. But the accuracy of the models is greater than 93%, so the effect is limited. Secondly, this paper only studies the total grain output at the national scale rather than specific to the provincial or country scale because of the limited length of the paper. Therefore, a future study can be conducted on the provincial or country scale so that more targeted suggestions can be given according to the different situation of each province or country.

This paper makes several contributions. Firstly, this paper suggests that it is necessary to apply scientific fertilization, increase agricultural water conservancy construction, and ensure farmland irrigation rate. In addition, pesticides should be used rationally, and efforts should be made to balance the distribution of agricultural mechanization to improve efficiency. Moreover, the plastic film should be used properly to ensure the robust growth of food crops, as well as natural disasters should be prevented and remedied. Secondly, this paper establishes a prediction model of grain yield to assist the food administrative department in making scientific and effective decisions. It is of great significance to national food security and grain production guidance.

Reference

Chen, Y. F, Li X, D. (2013). Spatial-temporal characteristics and influencing factors of grain yield change in China[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 29(20), pp.1—10.

Dobermann, A. (1994). Factors causing field variation of direct-seeded flooded rice. Geoderma, 62(1-3), pp.125-150.

Downing, T. E. (1992). Climate change and vulnerable places: global food security and country studies in Zimbabwe, Kenya, Senegal and Chile. Environment Change Unit[J], University of Oxford, 5(6), pp.28-30.

Einasto, M., Liivamägi, L. J., Saar, E., Einasto, J., Tempel, E., Tago, E., & Martínez, V. J. (2011). SDSS DR7 superclusters-Principal component analysis. Astronomy & Astrophysics, 535, A36.

Isik, M., & Devadoss, S. (2006). An analysis of the impact of climate change on crop yields and yield variability. Applied Economics, 38(7), pp.835-844.

Kanwar, S. (2006). Relative profitability, supply shifters and dynamic output response, in a developing economy. Journal of Policy Modeling, 28(1), pp.67-88.

Lin, L., & Su, S. (2004). Advances in the Study of Mineral Nutrients and Fertilization of Castanea mollissima BL. Journal of Beijing Agricultural College, 19(1), pp.73-76.

LIU, Y. (2010) Great Leap Forward and Chinese Famine of 1958—1961: State, Collective and Peasants in Centralized System, China Economic Quarterly, 9(3), pp.1119-1142.

Marín, S., Sanchis, V., Teixido, A., Saenz, R., Ramos, A. J., Vinas, I., & Magan, N. (1996). Water and temperature relations and microconidial germination of Fusarium moniliforme and Fusarium proliferatum from maize. Canadian journal of microbiology, 42(10), 1045-1050.

Mundry, R & Charles, N. (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. American Naturalist, 173(1), pp.119-123.

Samapundo, S., Devlieghere, F., De Meulenaer, B., Geeraerd, A. H., Van Impe, J. F., & Debevere, J. M. (2005). Predictive modeling of the individual and combined effect of water activity and temperature on the radial growth of Fusarium verticilliodes and F. proliferatum on

corn. International Journal of Food Microbiology, 105(1), pp.35-52.

Stepwise Regression in SPSS-Example. Retrieved December 10, 2018, from: http://www.spss-tutorials.com/stepwise-regression-in-spss-example/comment-page-1/

Timsina, J., & Connor, D. J. (2001). Productivity and management of rice–wheat cropping systems: issues and challenges. Field crops research, 69(2), pp.93-132.

Wheeler, T., & Von Braun, J. (2013). Climate change impacts on global food security. Science, 341(6145), pp.508-513.

Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). An overview of principal component analysis. Journal of Signal and Information Processing, 4(03), 173.

Jolliffe, I. (2011). Principal component analysis. In International encyclopedia of statistical science (pp. 1094-1096). Springer, Berlin, Heidelberg.