

# Research on Rice Yield Forecasting Model

Kui Fang<sup>1,\*</sup>, Qingshan Ren<sup>1</sup>, Xiangmei Feng<sup>2</sup> and Xinghui Zhu<sup>1</sup>

<sup>1</sup>College of Information Science and Technology, Hunan Agricultural University, Changsha, China

<sup>2</sup>Agricultural Information Center, Hunan Agricultural University, Changsha, China

\*Corresponding author

**Abstract**—The meteorological factors play an important role in rice yield. In this paper, according to the current agricultural meteorological factors on the impact of agricultural production, an the rice yield prediction model was established by using multiple stepwise regression analysis. The experimental results show that the average forecast accuracy is more than 98%, and the prediction result is consistent with the trend of the measured results, and the prediction results are credible.

**Keywords**—meteorological factor; rice yield forecasting model; correlation analysis; multiple stepwise regressions

## I. INTRODUCTION

The crop yield forecasting models mainly contain statistical forecasting model [1-2], remote sensing forecast model [3], dynamic (state) growth simulation forecast [4]. In this paper, we take Hunan early rice research as an example, and study early rice yield forecast statistical model.

The combined effects of meteorological factors such as temperature, light and water have important influence on the growth and development of rice [5].

In this paper, the data of early rice yield in Hunan (Hunan Provincial Statistical Yearbook) and the meteorological data of rice growth period (the basic data of Chinese agriculture in the Chinese meteorological data network) were selected as samples, we analyzed the correlation between meteorological yield and meteorological factors such as average temperature, late precipitation and sunshine hours during early growth period, and calculated the correlation coefficient between meteorological factors and meteorological yield.

Based on the physiological characteristics of early rice growth period, the key meteorological factors influencing the meteorological yield of early rice were selected, and the multiple regression model based on key meteorological factors was established by using multiple regression model. The model was used to forecast early rice yield.

## II. RICE YIELD FORECASTING MODEL

### A. Statistical Model of Crop Yield

There are many factors that affect the yield of crops. The crop yield can be decomposed into three parts: trend yield, meteorological yield and random yield. The statistical model of crop yield  $Y$  is:

$$Y = Y_t + Y_w + Y_\varepsilon \quad (1)$$

Where  $Y_t$  is the projected trend of production which is determined by social production conditions and social skill levels,  $Y_w$  is the predicted meteorological yield which represents the fluctuating part of the yield,  $Y_\varepsilon$  is a random yield, which is generated by accidental and statistical errors. In general,  $Y_\varepsilon$  is usually ignored. We sort the early rice yield data in Hunan during 1985~2010 by year (see Table I).

Suppose that  $x_i$  ( $i = 1985, 1986, \dots, 2011$ ) represents actual yield of year  $i$ , so we define

$$\text{Trend yield of year } i = (x_i + x_{i-1} + x_{i-2} + x_{i-3} + x_{i-4}) / 5 \quad (2)$$

This is 5-point moving average method.

At the same time, we define

$$\text{Meteorological yield} = \text{Actual yield} - \text{Trend yield} \quad (3)$$

Using above method, we show the trend yield and actual yield of rice in Hunan as shown in Table I.

TABLE I. TREND AND YIELD OF RICE IN HUNAN DURING 1985~2010 YIELD

Year	Actual yield	Trend yield	Meteorological yield
1985	5433.6		
1986	5715		
1987	5279.01		
1988	5476		
1989	5439.3	5468.582	-29.282
1990	5604.5	5502.762	101.738
1991	5280.3	5415.822	-135.522
1992	5261.9	5412.4	-150.5
1993	5103.1	5337.82	-234.72
1994	5530.1	5355.98	174.12
1995	5100.9	5255.26	-154.36
1996	5119.81	5223.162	-103.352
1997	5723.98	5315.578	408.402
1998	5158.69	5326.696	-168.006
1999	5203.33	5261.342	-58.012
2000	5789.68	5399.098	390.582
2001	5754.13	5525.962	228.168
2002	5126.91	5406.548	-279.638
2003	5294.76	5433.762	-139.002
2004	5560.82	5505.26	55.56
2005	5545.15	5456.354	88.796
2006	5513.68	5408.264	105.416
2007	5703.61	5523.604	180.006
2008	5919.98	5648.648	271.332
2009	5864.84	5709.452	155.388
2010	5619.26	5724.274	-105.014
2011	5779.4	5777.418	1.982

B. Correlation Analysis

The correlation analysis is used to analyze the degree of close relationship between variables. The size of the correlation coefficient indicates the degree of correlation between the variables. The magnitude of the correlation coefficient measures the magnitude of the interrelationship between the two variables. If the correlation coefficient between the two variables is large, we can say that there is a significant correlation between the two variables.

Because correlation analysis is the basis and prerequisite for regression analysis, multiple stepwise regression analysis based on correlation analysis is used. At the same time, regression analysis is the depth and continuation of correlation analysis. It is not enough to explain the close correlation between variables, and only the regression model can be used to make the expectation and estimation of the development trend of variables [6]. Therefore, the correlation between rice yield and meteorological factors was first analyzed.

We call the first third, or the middle third, or the last third of a month as a Xun. We know that Hunan early rice from March sowing to July mature harvest, a total of 12 Xuns. Let  $X_j^i$  be meteorological factors of a Xun of year  $i$ , where  $j=1,2,3$  respectively, mean average temperature, precipitation, sunshine hours. The  $Y_w^i$  represents the meteorological yield of year  $i$ .

For example, during 1989-2011, the average temperature  $X_1^i$  in early March, we obtain the correlation coefficient between variable  $X_1^i$  and variable  $Y_w^i$  is 0.394. And so on, we can calculate the correlation coefficient between meteorological factors  $Y_w^i$  and meteorological yield  $Y_w^i$  as shown in Table II.

TABLE II. CORRELATION COEFFICIENT TABLE

Time	Correlation coefficient		
	Average temperature	Precipitation	Sunshine hours.
Early March	0.394	-0.331	0.035
Late April	0.13	-0.272	-0.227
Mid- April	0.034	-0.032	0.044
Late April	-0.057	-0.368	0.004
Early May	0.6	-0.548	0.589
Mid-May	0.757	-0.412	0.499
Late May	0.341	-0.154	0.118
Early June	-0.2	0.222	-0.21
Mid-June	-0.068	-0.478	0.296
Late June	0.433	-0.543	0.524
Early July	0.613	-0.463	0.513
Mid- July	-0.098	-0.142	-0.208

According to the above Table II, all factors are related to meteorological production. Therefore, it can be seen that the correlation between meteorological factors and meteorological yield in early rice has obvious biological significance. So these factors can be used as key factor for the regression analysis of meteorological yield.

C. Multiple Regression Model for Forecasting Early Rice Yield

Based on the method of multiple stepwise regression method, the key meteorological factors were used as the independent variables and the meteorological yield as the dependent variables, and the relationship between the key meteorological factors and the meteorological yield was analyzed. Finally, the early rice yield forecast model was established as follows.

$$Y = b_0 + \sum_{i=1}^n b_i x_i \tag{4}$$

Where  $Y$  is the meteorological yield,  $x_i$  is the  $i$ th meteorological factor,  $b_0, b_1$  is pending coefficient.

In a multivariate stepwise regression, not all factors can enter the final model. In the process of establishing the model,

In the process of establishing the model, the model gradually filters the factor by F test, retains these factors that contribute greatly, and removes these factors with little contribution. Finally, the yield forecast model was obtained:

$$Y = -2270.972 + 7.474 * b_1 + 5.756 * b_2 - 1.526 * b_3 + 2.947 * b_4 + 1.423 * b_5 + 3.149 * b_6 - 0.896 * b_7 \tag{5}$$

In the forecasting model, only a total of 7 factors such as  $b_1, b_2, b_3, b_4, b_5, b_6, b_7$  are involved in modeling, and other factors with small influence factors are discarded. Where  $b_1$  is the

temperature in mid-May,  $b_2$  is the sunshine hours in mid-June,  $b_3$  is the precipitation in late April,  $b_4$  is the temperature in mid-May,  $b_5$  is the temperature in early April,  $b_6$  is the sunshine in late May Hours,  $b_7$  for precipitation in early June.

III. TEST OF MODEL

A. Test of Model for Fiting Goodness

In order to judge the degree of fitting of the regression line or curve to the sample value,the screening strategy of the stepwise regression is used to test the goodness of the regression equation, the correlation coefficient and its related indexes shown in Table III.

TABLE III. CORRELATION COEFFICIENT AND ITS RELATED INDEXES

Model	R(complex correlation coefficient)	R2 (judgment coefficient)	Adjusted R 2	Standard estimation error
1	0.978	0.957	0.932	52.96172

It can be seen from the above table that the standard estimation error is 52.96172,  $R = 0.978$ ,and the adjusted coefficient R2 is close to 1, which shows that the model has a high fitting goodness.

B. The Significance Test of the Model

1) *Test of linear relationships*: Using SPSS software, we can get the ANOVA table of the model discussed in this paper, as shown in Table IV.

TABLE IV. ANOVA TABLE

Model	Sum of squares	Degrees of freedom	Mean square	F	Statistical significance
7.000 Regression	750458.818	7.000	107208.403		
Residual	33659.326	12.000	2804.944	38.221	0.000
Total	784118.144	19.000			

It can be seen from the above table that the sums of squared deviations of the explanatory variables is 784118.144, the mean square is 107208.403, and the regression square sum is 750458.818. The residual sum of squares is 33659.326 and the mean square is 2804.944. The significance test of the regression equation shows that the test statistics  $F = 38.221$ , the corresponding probability is  $P = 0.000 < 0.05$  (significance level), so we can directly give a significance level  $\alpha = 0.05$ .

Because of  $K = 7, n - k - 1 = 14$ , By checking the F distribution table we get  $F_{\alpha}(k, n - k - 1) = F_{0.05}[7, 14] = 2.76$ .

Since  $F = 38.221 > 2.76$ , we can reject the  $H_0$  hypothesis, that the linear relationship of the model is considered to be significant at 95% level.

2) *Test of regression coefficient*: Using SPSS software, we can get the Coefficient table of the model discussed in this paper, as shown in Table V.

TABLE V. REGRESSION EQUATION COEFFICIENT TABLE

Model Project	Coefficient	Standard Error	t	Significance
(constant)	-2770.972	201.846	-13.728	0.000
Temperture in mid-May	7.474	0.772	9.678	0.000
Sunshine in mid-June	5.756	0.819	7.026	0.000
Precipitation in late April	-1.526	0.329	-4.643	0.001
Temperture in early May	2.947	0.720	4.093	0.001
Temperture in early April	1.423	0.542	2.626	0.022
Sunshine in late May	3.149	1.046	3.011	0.011
Precipitation in early June	-0.896	0.364	-2.459	0.030

From the above table we can see  $|t_1| = 9.678$ ,  $|t_2| = 7.026$ ,  $|t_3| = 4.643$ ,  $|t_4| = 4.093$ ,  $|t_5| = 2.626$ ,  $|t_6| = 3.011$ ,  $|t_7| = 2.459$ . Given a significant level  $\alpha = 0.05$ , because of  $K = 7, n - k - 1 = 14$ , by checking the t distribution table we obtain

$$t_{\alpha/2}(n - k - 1) = t_{0.025}(14) = 2.145$$

It can be seen that all of the  $|t|$  values by calculated are greater than the critical value, so reject the original hypothesis. That is, the seven explanatory variables, including the constant term, are significant at the 95% level and pass the variable significance test.

Based on the above analysis, the model passes the test of the linear relationship and the test of the regression coefficient. Therefore, the model can be used to forecast the early rice meteorological yield.

3) *Return test of model*: In order to test the forecasting effect of the model, the accuracy of the forecast is introduced, as follows.

$$\text{Accuracy} = [1 - (\text{simulated yield} - \text{actual yield}) / \text{actual yield}] \times 100\% \quad (6)$$

Through the forecast of early rice yield in 1990-2011 years, we see that the highest accuracy rate in 2011 and is 99.90%, the lowest accuracy rate in 2013, and is 98.31%, and the average accuracy of the forecast more than 98%.

IV. SUMMARY

For the early rice in Hunan Province, the meteorological yield and the meteorological factor of the main growth period were analyzed by using a multivariate stepwise regression analysis method based on correlation analysis, we can obtain these meteorological factors of significant correlation, which can use as the key meteorological factor, such as precipitation, sunshine and temperature, to establish the meteorological yield forecasting model of early rice. Further, the return simulation test of model shows the accuracy rate is not less than 98%. The model was tested on the meteorological factors and

meteorological yield data of 2012-2013 years early rice in Hunan, the accuracy of the forecast is more than 91%. It shows that this model can meet the needs of early rice agricultural meteorological yield forecasting and service.

#### ACKNOWLEDGMENT

This paper is supported by the research and development key projects in Hunan Province (No 2016NK2118) and national R & D key projects (No 2017YFD0301506).

#### REFERENCES

- [1] Shu-Yan L I, Liu W C. Research on the yield prediction of summer maize in Henan Province based on meteorological key factors[J]. Agricultural Research in the Arid Areas, 2014.
- [2] Liu W, Chen H, Yu W, et al. Dynamic Output Forecast Research for Winter Wheat Based on Climatic Suitability Index[J]. Meteorological & Environmental Sciences, 2008.
- [3] Huang J, Wu S, Liu X, et al. Regional winter wheat yield forecasting based on assimilation of remote sensing data and crop growth model with Ensemble Kalman method[J]. Transactions of the Chinese Society of Agricultural Engineering, 2012, 28(4):142-148.
- [4] Feng Y X, Zheng-Jin X U, Fan F, et al. Studies on Effects of Row Directions on Micrometeorological Characters of Rice (*Oryza sativa* L.) Population with Different Spike Pattern II. Effects of Row Direction on Dry Matter of Rice Population[J]. Agricultural Meteorology, 2004, 25(1):10-13.
- [5] Xie Na. Effect analysis on rice grain yield under climatic factor in Changsha county[D]. Hunan Agricultural University, 2010.
- [6] Cheng Zhengjiang, Pu Xi-an. A Comparative Study of Multiple Linear Regression Analysis and Stepwise Regression Analysis[J]. Mudanjiang College of Education, 2016, 05(171):131-133.
- [7] J. G. Jesse. The Elements of User Experience[M]. New Riders Press, 2007, 75-84.
- [8] C. Claudiu. PHP Language used for Faculty Site[J]. Journal of Computer Science and Control Systems, 2008, Vol.11, pp2-5.
- [9] G. Gupta, M.C. Govil. MVC Design Pattern for the multi framework distributed applications using XML, spring and struts framework[J]. International Journal on Computer Science and Engineering, 2010, Vol.24, p45-46.