Research on the prediction model of grain yield based on the ARIMA method

Fan Chao, Cao Pei-ge, Yang Tie-jun, Fu Hong-liang

College of Information Science and Technology, Henan University of Technology，Zhengzhou, China

Keywords: grain yield, forecasting, ARIMA model, prediction error

Abstract: In order to predict the grain yield of the country accurately, considering the periodical fluctuation of the data, the method of time series is used. Firstly, the stability and the relativity of the yield series from year 1980 to 2009 are analyzed, and the first-order difference of which is calculated to get a stationary series. Then, after comparing the value of AIC of different models, the forecasting model ARIMA(5,1,5) is selected as the best one, and the performance of which is tested. Lastly, the grain yields from year 2010 to 2012 are predicted by three different methods, the results shown that, the prediction error of the model ARIMA(5,1,5) is 4.478%, the error of the grey model GM(1,1) is 6.78%, and the error of the second exponential smoothing method is 7.682%, thus, the model ARIMA(5,1,5) is more suitable to forecast the grain yield in short-term.

Introduction

The amount of grain yield is increasing year by year and the food problem has been solved by China government, but there are many constraints, such as the decreasing of the farmland, the shortage of the water, the uneven distribution of the resource and so on. Thus, it is necessary and important to forecast the grain yield by analyzing the influence factors, which is helpful for the government to take actions to avoid the decreasing of the grain yield and guarantee the food security.

Presently, there are three main methods to estimate the yield of the grain, which is the meteorology forecasting method, the remote sensing technology and the statistical dynamics method, the prediction lead time for these methods is about only two months and the estimation error is about 5% to 10% of the total grain production, which can't need the requirement of long time estimation [1-6]. To resolve this problem, many other methods which can predict the grain yield one or several years ahead have been proposed, such as the multivariate regression model, the grey cognate analysis, the BP neural network and so on[7-9]. In fact, the production of the grain is affected by many factors, which induces the time series of the grain yield is non-stationary, and the autoregressive integrated moving average model (ARIMA) is just very suitable for the non-stationary time series and has been widely used.

Considering the characteristics of the grain production series and the merits of the ARIMA method, the optimal prediction model of the grain yield is established and the parameters of which is selected and tested by using the production data from year 1980 to 2012 of China.

The model of ARIMA

The ARIMA model was put forward by the statistician Box and Jenkins in early 70s of twenty

century, the basic principle of which is that, the data series which will be estimated is looked as a random series and is described by the mathematical model, the future value can be forecasted by using the present and the last data. The mode or structure of the series doesn't be supposed in advance, the data series can be described and fitted better by itself.

But in many situations, the time series is not stable, such as the data of GDP, the rate of exchange, the price of farm produce, and so on. For these non-stationary series, the series must be transferred into the stationary sequence by difference when using the ARMA method because the ARMA model is only suitable with the stable processing.

Supposed a non-stationary time series is denoted as {yi}, which can be transferred into a stationary sequence by d^{th} order difference, which is written as {xi}, where the series {xi} is a stable series, that is:

$$x_t = \Delta^d y_t \tag{1}$$

For the stationary series {xi}, the model of ARMA (p,q) can be established as:

$$x_t = c + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q} \tag{2}$$

Where xi (i=1,2,…,t) is the observed value at time i, p is the order of the autoregressive model, and φ1, φ2, ---, φp are the coefficients of AR; q is the order of the moving average model, and θ1, θ2,…, θq are the coefficients of MA; ut, ut-1,…,ut-q are the error at the time t, t-1,…, t-q respectively, c is a constant, and the c, φ1, φ2, ---, φp, and ut, ut-1,…,ut-q are the parameters of the ARMA model.

Introducing a delay operator L, that is Lxt=xt-1, thus, the order k delay operator can be defined as:

$$L^k x_t = x_{t-k} \tag{3}$$

The model of ARMA(p,q) can be written as:

$$(1 - \varphi_1 L - \varphi_2 L - \cdots - \varphi_p L^p) x_t = (1 + \theta_1 L + \theta_2 L + \cdots + \theta_q L^q) u_t + c \tag{4}$$

Or

$$\Phi(L) x_t = \Theta(L) u_t + c \tag{5}$$

Where, Φ(L)=1-φ1L-φ2L2- ---.-φpLp, which represents the autoregressive polynomials; Θ(L)=1+θ1L+θ2L+---.+θqL, which is the moving average polynomials.

Combining the Eq.(1) and (4), the ARIMA(p,d,q) model of non-stationary time series {yi} can be written as:

$$(1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p) \Delta^d y_t = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q) u_t + c \tag{6}$$

According to Eq.(6), we can see that, the essence of the ARIMA model is the combination of the difference with the ARMA model.

Construction of the forecasting model of the grain yield based on the ARIMA

The grain yield data of China from year 1980 to 2009 is used as the initial time series, which is denoted as X1 and shown in Fig.1. From these two figures, we can see that, the grain output

increases with the year, and the time series is a non-stationary sequence obviously. Thus, to predict the grain yield by the ARIMA model, the difference operation is used to the initial series to get a stationary sequence, which is denoted as X2 and the differenced series is shown in Fig.2.

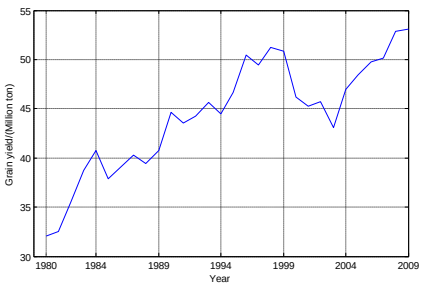

Fig.1. The grain yield from year 1980 to 2009

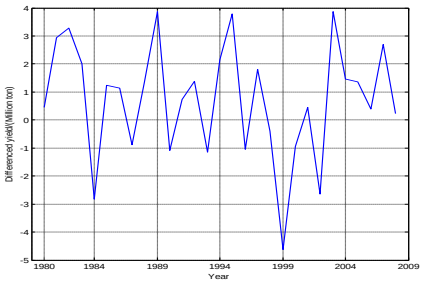

Fig.2. The difference of the grain yield

The ADF of the time series X1 and X2 is listed in table 1, for the initial series X1, the statistic value of t is -0.796466, which is larger than the testing threshold at the levels 1%，5%，10%, thus the hypothesis that the initial series is a non-stationary sequence is accepted. But for the differenced series X2, the value of t is -5.576412, which is less than the critical value at three testing levels, so, the suppose that the X2 is stationary series is true. then, the ARIMA forecasting model will be constructed.

Table 1. Results of ADF test for series X1 and X2

Series	Value of t	Threshold of t		
		Level 1%	Level 5%	Level 10%
X1	-0.796	-2.639	-1.952	-1.611
X2	-5.576	-2.642	-1.952	-1.610

For the ARIMA model, if the ARC and PARC are truncated after lag order p and q respectively, the p and q are just the order of the model. According to the ACF and the PACF of X2, we can see that, when p and q all less than five, the results are better. Thus, the AIC is calculated for $p \leq 5$ and $q \leq 5$, according to the results, when p and q are selected as the (5,1),(3,5) and (5,5), the values of AIC are relatively small, and the grain yields from year 2010 to 2012 are forecasted by using the models of ARIMA(5,1,1), ARIMA(3,1,5) and ARIMA(5,1,5) respectively, the predicted values and the actual amount of the grain are shown in table 2. According to the results, we can see that, the relative forecasting error of the model ARIMA(5,1,5) is least, thus, this model is selected as the optimal one, and the coefficients of the $\Phi(L)$ and $\Theta(L)$, which responds the Eq.(5), are written as:

$$\Phi(L) = 1 + 0.7953\,L + 0.1162L^2 - 0.02279L^3 + 0.07172L^4 + 0.5558L^5 \quad (8)$$

$$\Theta(L) = 1 + 1.226L + 1.089L^2 + 1.271L^3 + 0.4768L^4 + 0.01822L^5 \quad (9)$$

Table 2. The forecasting results for different model

		ARIMA(5,1,1)		ARIMA(3,1,5)		ARIMA(5,1,5)	
Year	Actual value	Predicted value	Error	Predicted value	Error	Predicted value	Error
2010	5.4641	5.4102	0.986%	5.3117	2.789%	5.3706	1.711%
2011	5.7121	5.3578	6.203%	5.2961	7.283%	5.3810	5.796%
2012	5.8957	5.4546	7.482%	5.3081	9.967%	5.5462	5.928%
Mean error			4.890%		6.679%		4.478%

Prediction of the grain yield

The outputs of the grain from year 2010 to 2012 are predicted by using the model ARIMA(5,1,5) based on the yield data from year 1980 to 2009, and the fitted error of the difference series X2 is shown in Fig.3. To compare the accuracy of this model, the outputs are calculated by using two other forecasting methods, which are the method of the grey forecasting model GM(1,1) and the method of second exponential smoothing, the prediction results are listed in table 3. According to the results, we can see that, in these three methods, the forecasting value of the model ARIMA is the most accurate.

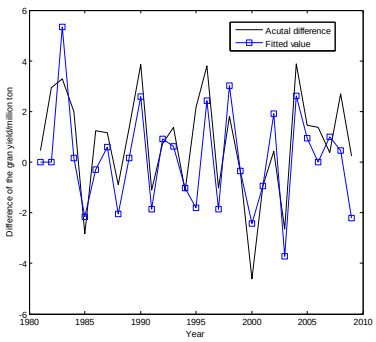

Figure 3. The fitted results of the time series of X2

Table 3. The comparison of three different forecasting methods

Year	Actual value	Predicted value		
		ARIMA(5,1,5)	GM(1,1)	Second exponential smoothing
2010	5.4641	5.3706	5.3554	5.1210
2011	5.7121	5.3810	5.4191	5.2142
2012	5.8957	5.5462	5.0812	5.4211
Mean error		4.478%	6.780%	7.682%

Conclusion

The prediction of the grain yield is a complex agricultural and statistical problem, which is affected many factors, such as the policy, the natural environment, the input of resources, and so on. Thus,

the output of the grain always changes abnormally with the variety of these elements, which induce the yields of the grain to increase and decrease alternately, and presents the periodical change. Considering the characteristics of the grain yield, the prediction model of ARIMA(5,1,5) is established, and the mean forecasting error of grain output for year 2010 to 2012 is only 4.478%, which is better than the grey model and the second exponential smoothing forecasting method, and it is very suitable to predict the grain yield accurately.

Acknowledgements

This project is supported by the science and technology support project of the State Grain Administration (No.201413001), and the State 863 Projects of China (No 2012AA101008).

Reference

1. Wang Hang, ZhuYan, Wenlong Li , Weixing Cao , Yongchao Tian. *Journal of Applied Remote Sensing*, v 8, n 1,pp:083674-1-6, January 2014, In Chinese
2. Weimin Ju, Ping Gao, Yanlian Zhou, *etal*. *International Journal of Remote Sensing*, v 31, n 6, p 1573-1587, February 2010 , In Chinese
3. A.D. Kleshchenko, T.A. Goncharova, T.A. Naidina. Using the satellite data in dynamic models of crop yield forecasting. *Russian Meteorology and Hydrology*, v 37, n 4, p 279-285, April 2012
4. Pute Wu, Xinxing Zhao. *Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering*, v 26, n 2, p 1-6, February 2010, In Chinese
5. O.D.Sirotenko, V.N Pavlova. A new approach to identifying the weather-crop yield functionals for assessing climate change consequences. *Russian Meteorology and Hydrology*, v 35, n 2, p 142-148, February 2010
6. D. Djigal, S.Saj, B. Rabary, E. Blanchart, *etal*. Mulch type affects soil biological functioning and crop yield of conservation agriculture systems in a long-term experiment in Madagascar. *Soil and Tillage Research*, v 118, p 11-21, January 2012
7. Liu Ran, Bu Hui. Study on nonlinear combination forecasting model for grain yield. *Information Technology Journal*, v 12, n 18, p 4666-4672, 2013
8. Zeren Chen, Haiyan Wang, Tingting Song, *etal*. The contrast of several ways in grain production prediction. *Journal of Convergence Information Technology*, v 6, n 12, p 248-256, December 2011
9. L.Naderloo, R. Alimardani, M. Omid, *etal*. Application of ANFIS to predict crop yield based on different energy inputs. *Measurement: Journal of the International Measurement Confederation*, v 45, n 6, p 1406-1413, July 2012