International Journal of Agriculture, Environment and Biotechnology
Citation: IJAEB: 8(2): 303-308 June 2015
DOI Number: 10.5958/2230-732X.2015.00037.6

Time Series Modeling for Trend Analysis and Forecasting Wheat Production of India

Ramesh Dasyam[1*], Soumen Pal[2], Vatluri Srinivasa Rao[3] and Banjul Bhattacharyya[1]

[1]Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, West Bengal-741252, India.
[2]Division of Computer Applications, ICAR- IASRI, New Delhi, India.
[3]Department of Statistics and Mathematics, ANGRAU, Bapatla, Andhra Pradesh, India.

[*]Corresponding author: dasyam.ramesh32@gmail.com

Paper No. 321 Received: 22 August 2014 Accepted: 26 April 2015 Published: 29 June 2015

Abstract

Wheat is one of the most important staple food grains of human for centuries. It has a special place in the Indian economy because of its significance in food security, trade and industry. This study made an attempt to model and forecast the production of wheat in India by using annual time series data from 1961-2013. Parametric regression, exponential smoothing and Auto Regressive Integrated Moving Average (ARIMA) models were employed and compared for finding out an appropriate econometric model to capture the trend of wheat production of the country. The best fitted model was selected based on the performance of several goodness of fit criteria viz. Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE), Akaike Information Criterion (AIC), Schwarz's Bayesian Information Criterion (SBC) and R-squared values. The assumptions of 'Independence' and 'Normality' of error terms were examined by using the 'Run-test' and 'Shapiro-Wilk test' respectively. This study found ARIMA (1,1,0) as most appropriate to model the wheat production of India. The forecasted value by using this model was obtained as 100.271 million tones (MT) by 2017-18.

Highlights

- Comparisons were made among parametric regression, Exponential smoothing (Holt) and ARIMA for selecting best fitted model.
- Superiority of ARIMA model was observed over the other models.

Keywords: Regression, normality, exponential smoothing, ARIMA

The share of wheat in the Indian foodgrain production is around 35.5% and it comprises about 22% of the total area under food grains. China and India are the top two wheat producing countries accounting for over 30% share in world production, however, the top two producers are also the major consumers of wheat and have a very small presence in the world trade. India has witnessed a substantial change in the past 4-5 decades with the overall wheat production increasing at a Compounded Annual Growth Rate (CAGR) of 4.22% during 1960-2010. The production of wheat in India at the same time, increased from 11 MT in 1960-61 to 80.8 MT in 2010-11. India produced close to a record 94.8 MT of wheat during 2012. The major wheat producing states in India are placed in the northern part of the country with Uttar Pradesh, Punjab and Haryana contributing to nearly 80% of the total wheat production. Wheat provides more nourishment for humans than any other food source. Globally, wheat is the leading source of vegetable protein in human food having higher protein content than maize or rice, the other major cereals. So, proper forecast is very important in an economic system for such an essential crop as it would be easier to formulate and initiate appropriate policy measures if data with regard to the trend of wheat production is obtained and analyzed in advance. Time series forecasting is an important statistical technique used as a basis for manual and automatic planning in many application domains (Gooijer and Hyndman 2006; Sonawane *et al.* 2013). In this present study, time series modeling was done for production of wheat in India by using parametric regression, exponential smoothing and ARIMA models and out of sample accuracy for each of the models has been computed. As the ARIMA model outperforms other methods for this particular dataset, the final forecasting of wheat production for the year 2014-15 till 2017-18 has been estimated by using this approach.

Materials and Methods

Data with respect to wheat production of India for period of 1961-2013 was collected from Directorate of Economics and Statistics, Department of Agriculture and Cooperation, Government of India. In this, last three years data were used for model validation and remaining for model building. Before analysis, as the study is dealing with time series, present data set have been verified initially for existence of outlier.

Test for Outlier

For detecting the outlier in the time series, Grubbs test was used in the current scenario as the test is particularly useful in case of large sample and easy to follow. Graph pad software which is widely used, has been employed in the present study to identify the existence of outliers and if found, have been replaced by the median of respective series (Sahu 2010). The statistic Z is calculated as absolute value of difference between the observation and mean divided by the Standard Deviation (SD) as shown in Eq. 1:

$$Z = \frac{\left|\bar{X} - X_i\right|}{SD_X} \qquad (1)$$

Here SD_x is the Standard deviation of variable X and i indicates the number of observation. If Z is greater than 1.96, then it is implied that outlier exists in the present sample.

ARIMA Model

According to Box and Jenkins (1976), a non seasonal ARIMA model is denoted by ARIMA (p,d,q) which is a combination of Auto Regressive (AR) and Moving Average (MA) with an order of integration or differencing (d), where p and q are the order of autocorrelation and moving average respectively (Gujarati *et al.* 2012).

The Auto-regressive model of order p denoted by AR(p) is as follows:

$$Z_t = c + Ø_1 Z_{t-1} + Ø_2 Z_{t-2} + \ldots + Ø_p Z_{t-p} + e_t \qquad (2)$$

where c is constant term, $\boldsymbol{Ø}_p$ is the p-th autoregressive parameter and e_t is the error term at time t.

The general Moving Average (MA) model of order q or MA(q) can be written as:

$$Z_t = c + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots\ldots - \theta_q e_{t-q} \qquad (3)$$

where c is constant term, $\boldsymbol{\theta}_q$ is the q-th moving average parameter and e_{t-k} is the error term at time t-k.

ARIMA in general form is as follows:

$$\boldsymbol{\Delta}^d Z_t = c + (\boldsymbol{Ø}_1 \Delta^d Z_{t-1} + \ldots + Ø_p \boldsymbol{\Delta}^d Z_{t-p}) - (\boldsymbol{\theta}_1 e_{t-1} + \ldots + \theta_q e_{t-q}) + e_t \qquad (4)$$

where $\boldsymbol{\Delta}$ denotes difference operator like

$$\boldsymbol{\Delta Z}_t = Z_t - Z_{t-1} \qquad (5)$$

$$\boldsymbol{\Delta}^2 Z_{t-1} = \Delta Z_t - \Delta Z_{t-1} \qquad (6)$$

Here, Z_{t-1}, …, Z_{t-p} are values of past series with lag 1, …, p respectively.

Modeling using ARIMA methodology consists of four steps viz. model identification, model estimation, diagnostic checking and forecasting (Sankar 2011).

Model identification and estimation

Model identification by ARIMA (p, d, q) is based on the concept of time-domain analysis i.e. autocorrelation function (ACF) and partial autocorrelation function (PACF). The ACF and PACF play vital role for the internal structure of the analyzed series. The ACF at lag k of the y_t series denotes the linear correlation coefficient between y_t and y_{t-k} , calculated for k = 0, 1, 2, 3,.. and so on as shown in Eq. 7. The PACF was calculated as the linear correlation between y_t and y_{t-k}, controlling for possible effects of linear relationships among values at intermediate lags.

$$\rho_k = \frac{\operatorname{cov}(y_t, y_{t-k})}{\sqrt{\operatorname{var}(y_t)\operatorname{var}(y_{t-k})}} \quad ....(7)$$

In this present study, Augmented Dickey Fuller (ADF) test has been used to find unit root in the time series data of variable under consideration (Dickey and Fuller 1979). For identification of data stationarity, line graph has been applied to represent the graphical behavior of observation at level, first difference and so on (Gaynor and Kirkpatrick 1994). Once the order of differencing has been diagnosed, the differenced univariate time series can be analyzed by the method of time-domain.

After identification of the appropriate p and q values for the model, the parameter of the autoregressive and moving average terms have been estimated. Standard statistical package SAS was used to estimate relevant parameters using iterative procedure.

Diagnostic checking

The estimated model was checked to verify if it adequately represents the series or not further. For evaluating the adequacy of ARIMA process, various reliability statistics have been used. Diagnostic checks including investigation of residual plots for ACF and PACF, Histogram-Normality and Randomness tests of residuals i.e., Shapiro-Wilk and Run tests were applied in the present study. The model with minimum values of RMSE, MAPE, MAE, MSE, AIC, SBC and with high R-squared value was considered as an appropriate model for forecasting (Shafaqat 2012).

Parametric Regression model

Other than ARIMA model, parametric regression models like Linear, Quadratic, Exponential, Power and Logarithmic models have been applied for modeling of wheat production. The models are given by Eq. 8 through Eq. 12:

(i) Linear: $Z_t = a + bt + e_t$ (8)

(ii) Quadratic: $Z_t = a + bt + ct^2 + e_t$ (9)

(iii) Exponential: $Z_t = a \operatorname{Exp}(bt) + e_t$ (10)

(iv) Power: $Z_t = at^b + e_t$ (11)

(v) Logarithmic: $Z_t = a + b \ln(t) + e_t$ (12)

where a, b, t and e_t represent constant, regression coefficient, time and error term respectively in the models.

Exponential smoothing

In addition to above models, Holt (Double exponential smoothing) method has been employed for modeling of non-seasonal time series wheat production data with trends. The model is expressed by two equations to deal with one for Level (α) and other for Trend (β) as shown in Eq. 13 and 14 respectively. α and β can assume values from 0 to 1 whereas optimum values of these two parameters have been estimated by minimizing the MSE over observations of data set.

$$L_t = \alpha y_t + (1-\alpha)(L_{t-1} + b_{t-1}) \quad (13)$$

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \quad (14)$$

where L_t denotes estimate of the level and b_t is the trend (slope) of the series at time t.

Results and Discussion

At first, wheat production data in India from 1961-2010 was tested for outliers by Grubbs method. It was observed that the number of extreme observations in the present data was zero, as presented in Table 1.

Table 1. Grubbs test for detecting Outliers

Mean:	44.6748
SD:	23.0635
No of observations:	50
Outlier detected?	No

Before analyzing by ARIMA, parametric regression and Holt models were applied to the dataset under consideration. From Table 2, it can be concluded that the Quadratic model was superior to other selected regression models based on diagnostic criteria. It might be due to time series data of wheat production follows quadratic growth pattern.

Similarly, parameters of Holt model were estimated as level (α) = 0.539 and trend (β) = 0.001 and depicted in Table 3.

After consideration of these models viz. Quadratic and Holt, ARIMA technique was employed in addition. At first, stationarity of wheat production in India from 1961-2010 was tested by time series plots and ADF test. The time series plot clearly indicated that the data was non stationary because of prominent increasing trend as shown in Figure 1.

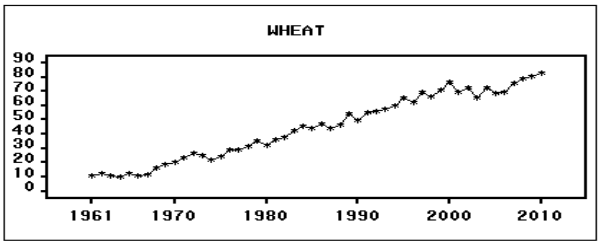

Figure 1. Time series plot of wheat production

ADF test for unit root also confirmed that the data was nonstationary and it became stationary at first difference as the calculated values were lesser than critical values at 1%, 5% and 10% levels (Table 4). It is also clear from the trend of time series plot at first difference as revealed in Figure 2.

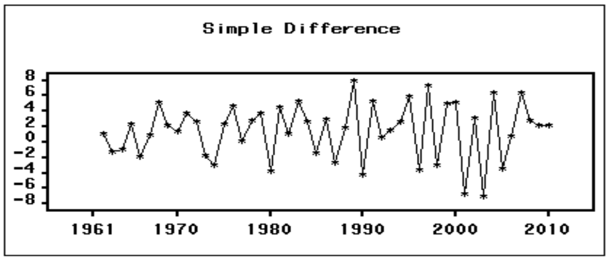

Figure 2. Time series plot for first differenced wheat production

After fixing the value of d as 1, values of p and q were determined. From correlogram of ACF and PACF as shown in Figure 3, there was only one significant spike for both ACF and PACF at lag 1.

Table 2. Parametric regression models for estimation of wheat production

Model	R^2	RMSE	MAPE	MAE	MSE	Fitted Equation
Linear	0.968	3.381	8.653	2.625	11.433	Z_t= 4.775 + 1.564t + e_t
Quadratic	0.978	3.361	8.403	2.595	11.294	Z_t= 83.691 + 1.462t - 0.002t^2 + e_t
Exponential	0.826	9.513	16.642	6.915	90.512	Z_t= 2.525 Exp (0.043t) + e_t
Power	0.948	5.191	14.209	4.318	26.946	Z_t = 1.544 $t^{0.699}$ + e_t
Logarithmic	0.791	10.435	32.934	8.624	108.908	Z_t = -23.834 + 23.071 ln(t) + e_t

Table 3. Exponential Smoothing models for estimation of wheat production

Model	R^2	RMSE	MAPE	MAE	MSE	Estimation of Parameters
Holt	0.980	3.142	7.997	2.769	9.873	α= 0.539, β=0.001

Table 4. Result of ADF test

Test	ADF statistic	Critical values at			Prob.	Decision
		1%	5%	10%		
ADF at level	-2.978	-4.161	-3.506	-3.183	0.148	Data Non-Stationary
ADF at first difference	-11.016	-4.161	-3.506	-3.183	0.0001	Data Stationary

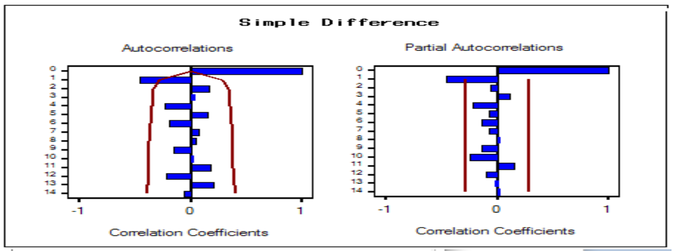

Figure 3 Correlogram of ACF and PACF for first differenced wheat production

In this present work, possible ARIMA (p,d,q) models such as (1,1,1), (0,1,1) and (1,1,0) were compared to each other. Among all possible models, ARIMA (1,1,0) was selected as optimal and most appropriate model due to model selection criteria such as minimum values of RMSE, MAPE, MAE, MSE, AIC, SBC and high R-squared value (Table 5).

Table 5. ARIMA Model Fit statistics

Model	R-squared	RMSE	MAPE	MAE	MSE	AIC	SBC
(1,1,1)	0.980	3.141	7.866	2.585	9.861	118.027	123.702
(0,1,1)	0.978	3.172	8.002	2.631	10.049	117.069	120.852
(1,1,0)	0.981	3.136	7.791	2.576	9.838	116.145	119.928

It was found that ARIMA model performed better than the earlier selected models viz. Quadratic and Holt. The parameters were estimated for the best selected model i.e., ARIMA (1,1,0) as mentioned in Table 6.

Table 6. ARIMA (1,1,0) Model Parameters estimation

Model Parameter	Estimate	Std. Error	t	Sig.
Intercept	1.46373	0.3177	4.6074	0.0001
Autoregressive, Lag 1	-0.44770	0.1295	-3.4562	0.0012

From Table 6, equation of the ARIMA model was formulated as: Wheat production$_t$ (Z_t) = 1.4637 – 0.4477 Z_{t-1} + e_t

From the residual ACF and PACF plots of ARIMA (1,1,0), it was clear that all autocorrelations and partial autocorrelations lie between 95% control limits as shown in Figure 4. This also confirmed the 'good fit' of this selected model.

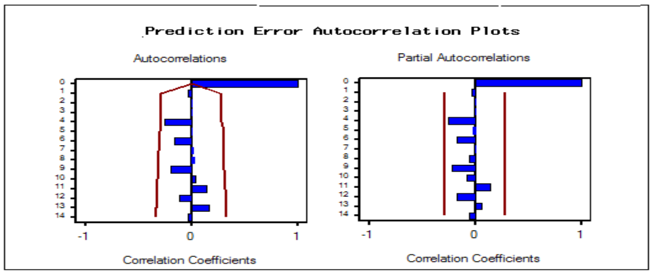

Figure 4 Residual ACF and PACF of ARIMA (1,1,0)

For checking normality and randomness, Shapiro-Wilk and Run tests were applied respectively to residuals of ARIMA(1,1,0) and results were presented in Table 7. The probability values for the both the tests were greater than 0.05 indicating residuals were distributed normally and independently. Histogram of residuals is depicted in Figure 5 which further confirmed the normality for the residuals.

Table 7. Tests of Normality and Randomness of residuals

	Shapiro-Wilk			Run test		
	Statistic	df	Sig.	Z-value	No of Runs	Sig.
Residuals	0.984	49	0.741	-0.801	22	0.423

Finally, forecasting was done for wheat production of India from 2011-12 till 2017-18 by using ARIMA (1,1,0) with keeping first three years data for validation. Predicted values with 95% Upper control limits (UCL) and Lower control limits (LCL) were presented in Table 8.

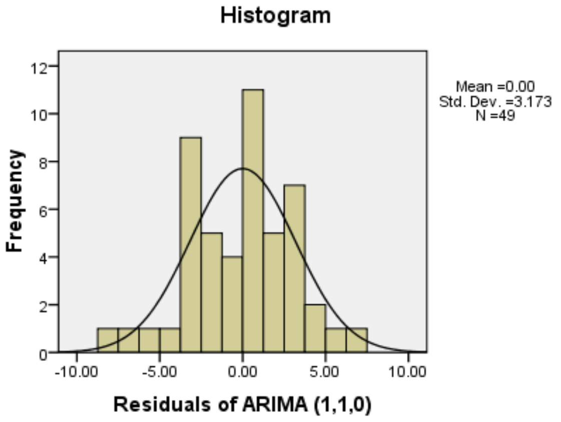

Figure 5. Histogram for Residuals of ARIMA (1,1,0)

Table 8. Forecasting of Wheat production with control limits

Year	Predicted (MT)	UCL (MT)	LCL (MT)	Actual (MT)	Absolute Forecast Error
2011-12	88.414	97.059	83.769	86.874	0.0177
2012-13	92.696	100.179	84.613	94.880	0.0230
2013-14	93.843	103.128	84.557	93.510	0.0035
2014-15	95.498	105.891	85.105		
2015-16	97.072	108.530	85.614		
2016-17	98.678	111.086	86.269		
2017-18	100.271	113.571	86.971		

By using ARIMA (1,1,0), it was observed that the actual and predicted values were closely related and predicted values were within control limits as captured in Figure 6.

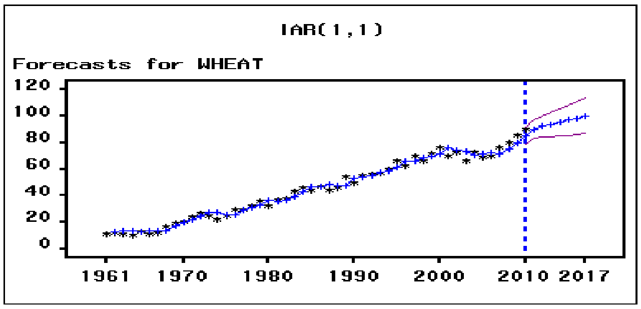

Figure 6. Forecasting of wheat production by ARIMA (1,1,0)

Conclusion

Present study made an attempt towards short term prediction of wheat production in India for upcoming years. Box and Jenkins methodology of univariate ARIMA model has been selected as an appropriate econometric model than traditional parametric regression and Holt smoothing models. ARIMA (1,1,0) was found as most appropriate among other ARIMA models and hence employed in forecasting wheat production of India. From the forecasted values, it can be concluded that for a few coming years production of wheat will follow an increasing trend and it has been estimated as 100.271 MT for the year 2017-18. These projections can play vital role to deal with future food security measures and planning for policy makers in India. Finally, increasing agriculture funding, selection of high yielding varieties and enhancing relationship between farmers and research workers may be important factors in sustaining this trend of production for long term.

References

Box GE and Jenkins GM 1976. Time Series Analysis: Forecasting and Control. Holden Day, San Francisco.

Dickey DA and Fuller WA 1979. Distribution of estimators for Autoregressive Time Series with a Unitroot. *Journal of the American Statistical Association* **74**: 427-431. doi: 10.1080/01621459.1979.10482531

Gaynor PE and Kirkpatrick RC 1994. Introduction to time series modelling and forecasting in business and economics. Mc Graw Hill, New York.

Gooijer JGD and Hyndman RJ 2006. 25 years of time series forecasting. *International Journal of Forecasting* 22: 443–473. doi:10.1016/j.ijforecast.2006.01.001

Gujarati D, Porter D and Gunasekar S 2012. Basic Econometrics. Mc Graw Hill, New Delhi.

Sahu PK 2010. Forecasting production of major food crops in four major SAARC countries. *International Journal of Agricultural and Statistics Science* **10**:71-92.

Sankar J 2011. Forecasting fish product export in Tamilnadu - A stochastic model approach. *Recent Research in Science and Technology* **3**:104-108.

Shafaqat M 2012. Forecasting Pakistan exports to SAARC - An application of univariate ARIMA model. *Journal of Contemporary Issues in Business Research* **1**:41-54.

Sonawane A, Hasan M, Rajwade Y, Desai S, Rajurkar G, Shinde V, Singh D and Singh M 2013. Comparison of Neuro-Fuzzy and Regression Models for Prediction of Outflow of on-farm Reservoir. *International Journal of Agriculture, Environment and Biotechnology* **6**:311-316.