



ARIMA Vs VARMA - Modelling and Forecasting of India's Cereal Production

S. Ravichandran and B.S. Yashavanth

ICAR-National Academy of Agricultural Research Management, Hyderabad

Received 27 April 2019; Revised 24 January 2020; Accepted 14 February 2020

SUMMARY

In agriculture, data on various parameters such as area, production and yield are collected over time. These data collected over time are modelled using various time-series modelling techniques. In this paper, an attempt is made to model time-series data of two important food commodities viz. Rice and Wheat using Autoregressive Integrated Moving Average (ARIMA) model and its multivariate variant Vector Autoregressive Integrated Moving Average (VARMA) model. The VARMA models are advantageous over the ARIMA models since two or more series can be modelled simultaneously besides capturing the relations between different series. The performance of ARIMA and VARMA models are compared using the measures of accuracy. Time-series data on production of rice and wheat for the period 1965-2017 is utilized for modelling and forecasting using ARIMA and VARMA statistical time-series modelling techniques. It was observed that the multivariate VARMA modelling technique is not an alternative to the univariate ARIMA modelling technique in terms of efficiency since the production of these two commodities are independent of each other. Finally, forecasting of rice and wheat production for the year 2020 was carried out and is found out to be 114 million tonnes of rice and 106 million tonnes of wheat. An increase of 4.5 % in rice production and 8.8 % in wheat production over the current production values are forecasted for the year 2020. Forecasting for future years is essential as this would help the planners in planning for eventualities arising due to vagaries of monsoon such as floods or droughts.

Keywords: Time-series, ARIMA, VARMA, Forecasting, India.

1. INTRODUCTION

India lives in villages. Sixty-five percent of Indian population lives in villages for whom agriculture is the main source of income. Agricultural growth rate was pegged at 4% in 2016-17. Rice and Wheat are the major cereal crops grown in the Indian subcontinent. Together, they were grown in more than 72 million hectares in 2012-13. Globally, India stands second in production of rice with nearly 42 million hectares under rice cultivation producing around 106 million tonnes. Similarly, India also stands second in wheat production with nearly 30 million hectares under cultivation producing 95 million tonnes of wheat. This high production values have made India self-sufficient in these two cereals. Though the production of rice and wheat have an increasing trend over the years, it is necessary to forecast their production using sound

statistical modelling techniques. These forecasts will be useful to the governments and agribusiness industries to execute policies for providing technical and market supports.

In statistical time-series modelling, the historical data of the variables under study are collected, analyzed and subsequently the models describing the underlying relationship are developed. In recent past, much effort has been directed towards developing and improving time-series forecasting models. The most important and widely used time-series model is the Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) modelling and forecasting methodology. The ARIMA model is popular among the researchers because of its statistical properties. The applicability of well-known Box-Jenkins methodology in the model building process (Box *et al.*, 1994) makes it an appropriate choice for

modelling and forecasting. Many researchers have used ARIMA technique to study the production of several agricultural commodities like Rice (Raghavender, 2010; Rahman, 2010; Suleman and Sarpang, 2011, Ravichandran, 2012; Jhambulkar, 2013), Maize (Badmus and Ariyo, 2011) and Wheat (Iqbal et al., 2005). However, when more than one variable is to be modelled, ARIMA necessitates each series to be modelled separately consuming good amount of time and other resources. This problem can be addressed by applying multivariate time-series analysis such as Vector Autoregressive Moving Average (VARMA) models (Lütkepohl, 2006). In VARMA models, all the series under study are modelled together simultaneously. Besides, the VARMA modelling method is simple since all variables considered are endogenous and the Ordinary Least Square technique (OLS) can be applied for estimation of parameters making it advantageous over other multivariate modelling techniques like simultaneous equation models (Gujarati *et al.*, 2009). Also, since all the variables are considered simultaneously, it helps in capturing the relations between different series giving us better models than those by the univariate time-series models. Studies have been attempted to use Vector Autoregression (VAR) models to study fish landings (Sathianandan, 2007), oil prices (Kilian, 2011) and coffee prices (Yashavanth *et al.*, 2017). The present study focuses on the application of VARMA time-series model to model and forecast the annual production of Rice and Wheat in India and compare the results obtained with the ARIMA time-series modelling procedure.

2. MATERIALS AND METHODS

2.1 Data

Time-series data on annual production of rice and wheat in India from the year 1965 to 2017 available at www.indiastat.com was utilized for the present study. As per time-series modelling procedure, part of the data is utilized for developing appropriate model and part for validating the developed model. Accordingly, in the present study, data pertaining to 1965 to 2012 are utilized for developing appropriate model and the remaining data from 2013 to 2017 are utilized to test the developed model.

2.2 ARIMA Approach

ARIMA time-series modelling procedure due to Box *et al.* (1970) are an extension to the ARMA

(Autoregressive Moving Average) models (Whittle, 1951) and comprise the most general class of time-series models useful in modelling and forecasting non-stationary time-series data. In an ARIMA model, it is assumed that the time-series variable under study is a linear function of the past actual values and random shocks.

In general, an ARIMA model, represented as ARIMA (p, d, q), comprises three components: (i) p , the order of Auto-Regression (AR); (ii) d , the order of integration (differencing) to achieve stationarity; and (iii) q , the order of Moving Average (MA). ARIMA technique is a parsimonious approach representing both stationary and non-stationary processes.

Consider an ARMA (p, q) process defined by Equation (1)

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \quad (1)$$

where, y_t are the actual values and ε_t are random shocks at time period t . ϕ_i ($i=1, 2, \dots, p$) and θ_j ($j=1, 2, \dots, q$) are the model parameters. The random errors, ε_t are assumed to be independently and identically distributed with mean of zero and constant variance, σ^2 . The first and most important requirement in ARIMA modelling is to ensure that the series under study is stationary since the estimation procedure is available only for a stationary series. A series is regarded stationary if its statistical characteristics such as the mean and the autocorrelation structures do not change over time. The stationarity of a time-series can be confirmed either by a time plot, which requires some experience or by using unit root tests like the parametric Augmented Dicky-Fuller test (Dickey and Fuller, 1979) or the non-parametric Phillips-Perron unit root test (Phillips and Perron, 1988). If a series is found to be non-stationary based on these tests, it can be made stationary by differencing. The number of times a series is differenced to achieve stationarity is referred to as the order of integration, d . Sometimes, other appropriate transformations like logarithmic transformation are used to achieve stationarity. Once the stationarity of the series is established, the Box-Jenkins approach, which goes in four steps viz., (i) identification (ii) estimation (iii) diagnostic checking and (iv) forecasting is employed. In the identification stage, multiple ARMA models with different values for q (MA terms) and p (AR terms) are chosen based

on Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF), respectively. This is followed by the estimation stage, where parameters of the tentative models are estimated by employing any of the non-linear optimization procedures such that the overall measure of errors is minimized or the likelihood function is maximized. Among all the candidate models, the best suited ARIMA model is selected by using Information Criteria. The model which has the smallest Akaike Information Criterion (AIC) (Akaike, 1974) or Schwarz-Bayesian Criterion (SBC) (Schwarz, 1978) value is chosen as the best suited model for the data under study. In the diagnostic checking stage, the residuals are checked for the normality and adequacy of the model. In the final forecasting stage, the future values are forecasted using the chosen model.

2.3 VARMA Model

The ARIMA model involves only one variable. In an ARIMA (p, d, q) , we regress a variable on p lags of itself and q lags of the random shocks. In contrast, a vector autoregressive moving average model, or VARMA (p, q) , involves k variables and k different equations are estimated. In each equation, we regress the relevant left hand-side variable on p lags of itself and all other variables, q lags of shocks of itself and all other variables. Thus the right hand side variables are the same in every equation – p lags of every variable and q lags of every random shock (Tiao and Tsay, 1989; Lütkepohl, 2005). This allows for cross-variable dynamics in VARMA models, which is absent in univariate models.

For k time-series $\{Y_{1t}\}, \{Y_{2t}\}, \dots, \{Y_{kt}\}$, $(t=0, 1, 2, 3, \dots, n)$ at equally spaced time intervals, the components can be represented by a vector $\mathbf{Y}_t = (\mathbf{Y}_{1t}, \mathbf{Y}_{2t}, \dots, \mathbf{Y}_{kt})^T$ called as a vector of time-series. A vector time-series with k components can be modelled by a VARMA model of order (p, q) denoted by VARMA (p, q) , and its expression is

$$\mathbf{Y}_t = \boldsymbol{\mu} + \beta_1 \mathbf{Y}_{t-1} + \beta_2 \mathbf{Y}_{t-2} + \dots + \beta_p \mathbf{Y}_{t-p} + \theta_1 \boldsymbol{\varepsilon}_{t-1} + \theta_2 \boldsymbol{\varepsilon}_{t-2} + \dots + \theta_q \boldsymbol{\varepsilon}_{t-q} + \boldsymbol{\xi}_t \quad (2)$$

where $\boldsymbol{\mu}$ is the mean vector of the series, β_i ($i=1, 2, \dots, p$) and θ_j ($j=1, 2, \dots, q$) are $k \times k$ parameter matrices, $\boldsymbol{\xi}_t = (\xi_{1t}, \dots, \xi_{kt})^T$ are independently and identically distributed random innovation vectors with zero mean and constant dispersion matrix Σ .

2.4 Forecast Accuracy Measures

The ability of different models to forecast the time-series values is assessed by using two common performance measures, viz. the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE) (Hyndman and Koehler, 2006). The RMSE measures the overall performance of a model and is given by Equation (3)

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (\hat{y}_t - y_t)^2} \quad (3)$$

where, y_t is the actual value for time t , \hat{y}_t is the predicted value for time t , and n is the number of predictions. The second criterion, the MAPE is a measure of average error for each point forecast and is given by Equation (4)

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{\hat{y}_t - y_t}{y_t} \right| \times 100 \quad (4)$$

where, the symbols have the same meaning as above. The model with least RMSE and MAPE values is considered as the best model for the data.

3. RESULTS AND DISCUSSION

3.1 Descriptive Analysis

The Table 1 gives the descriptive statistics of the production data for rice and wheat used in the study. The production has increased continuously over the time and reported the maximum production of 109.1 million tonnes and 97.4 million tonnes in 2016-17 for rice and wheat respectively. The measures of dispersion were found to be almost same for both the series.

Table 1. Descriptive statistics of production data used in the study (in million tonnes)

Series	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
Rice	30.4	109.1	70.2	23.5	-0.021	-1.286
Wheat	10.4	97.4	54.3	24.8	-0.019	-1.118

The foremost step in time-series analysis is to plot the data under study to have a visual inspection over its behaviour. Fig. 1 shows the time-series plot of annual production of rice and wheat in India for the periods 1965-2017.

A perusal of Fig. 1 shows a positive trend in production values indicating non-stationary of both the series. The Augmented Dickey-Fuller (ADF) test is performed to confirm the presence of non-stationarity

Production of Wheat & Rice

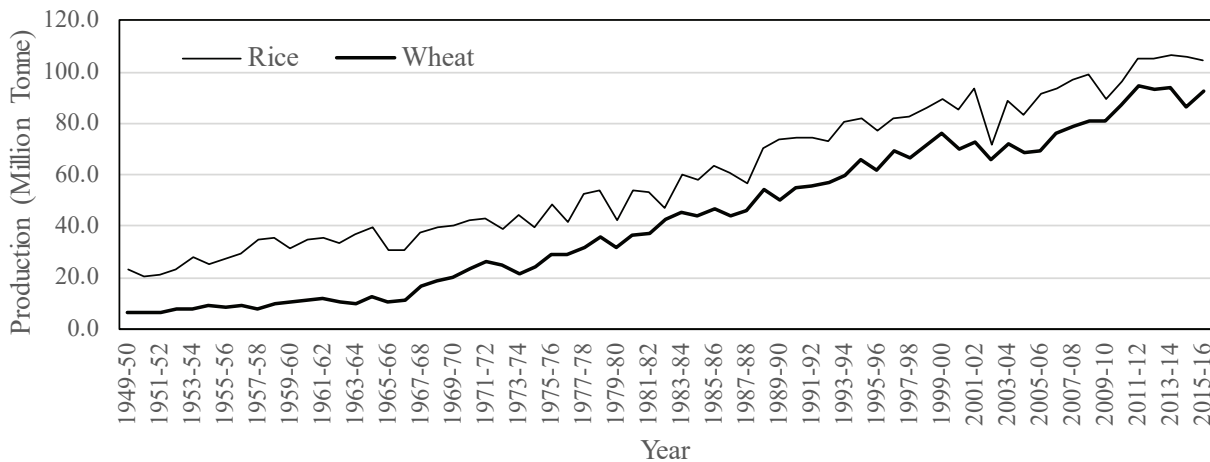


Fig. 1. Trend in production of rice and wheat in India

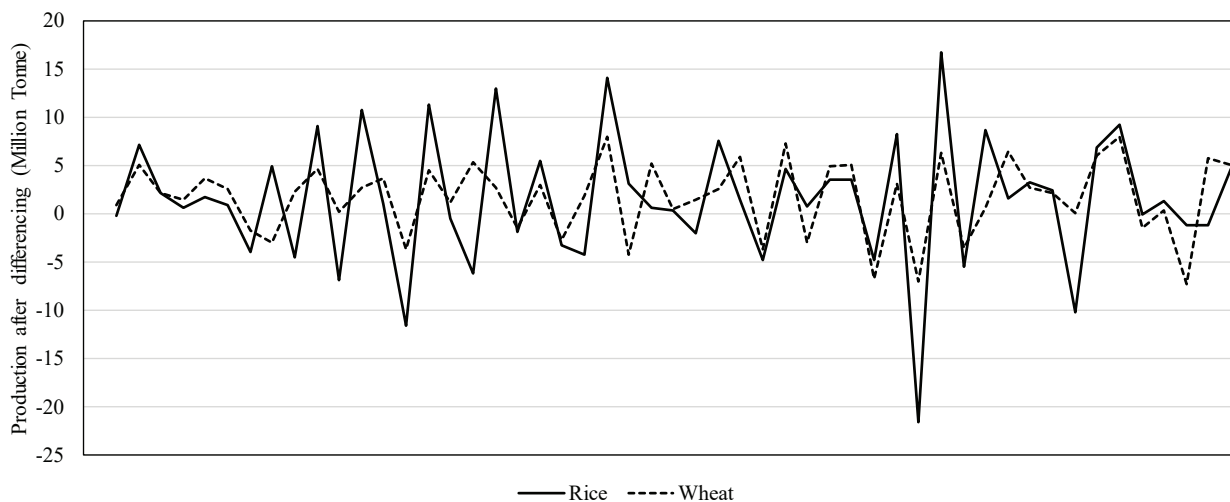
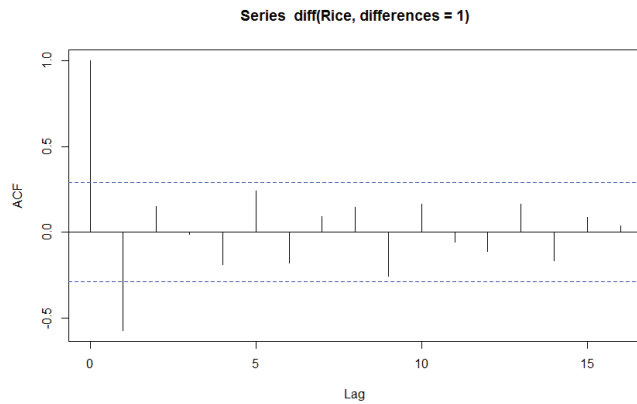


Fig. 2. Annual production of rice and wheat in India, after first differencing

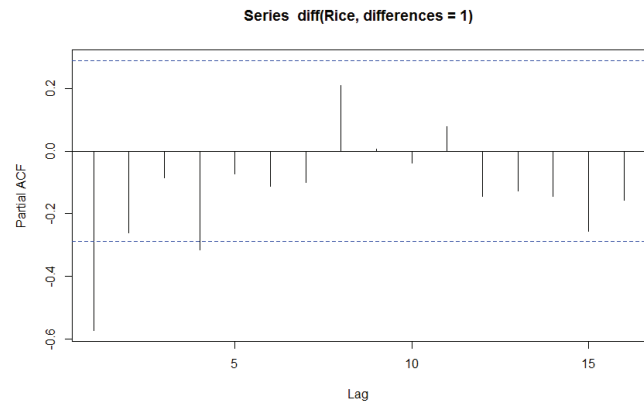
in the original data, results of which are given in Table 2. The Table also includes the results of ADF test performed after first differencing. The values clearly indicate the non-stationarity of both original series. Hence, the series were differenced and the ADF test results indicate that the differenced series are stationary and no further differencing is needed. Fig. 2 gives the plot of annual production of rice and wheat after first differencing. Unlike as in Fig. 1, the differenced series plotted in Fig. 2 does not show any upward trend and the fluctuations are also more or less same throughout. Thus, it is evident from the figure that the mean and variance are more-or-less constant throughout indicating the stationarity of the series.

Table 2. Augmented Dickey-Fuller test for stationarity

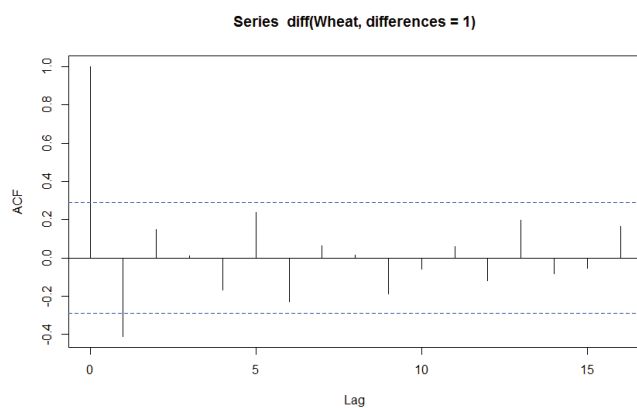
Series	Type	Original series		1st differenced series	
		ADF value	p-value	ADF value	p-value
Rice	Zero Mean	0.92	0.891	-88.29	0.0001
	Single Mean	-0.84	0.896	-136.36	0.0001
	Trend	-30.38	0.002	-136.57	0.0001
Wheat	Zero Mean	1.38	0.952	-27.70	0.0001
	Single Mean	-0.05	0.951	-72.16	0.0003
	Trend	-16.35	0.099	-71.13	0.0001



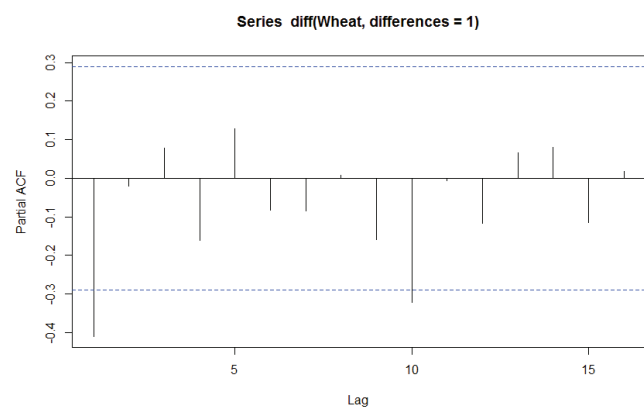
(a) ACF for differenced rice series



(b) PACF for differenced rice series



(c) ACF for differenced wheat series



(d) PACF for differenced wheat series

Fig. 3(a)-3(d). ACF and PACF values of the differenced series of Rice and Wheat

Once we achieved stationary time-series, the autocorrelation function (ACF), partial autocorrelation function (PACF) are used to find out the ARMA structures. The ACF and PACF of the original as well as the differenced series are plotted in Fig. 3(a)-3(d). The Fig. 3(a) and 3(c) give the values of auto-correlation of differenced rice and wheat series with their respective lagged values up to 16 lags. Similarly, Fig. 3(b) and 3(d) give the correlation of the differenced rice and wheat series with their residuals. It is evident from these plots that the values for ACF are significant at lags 0, 1 and values for PACF are significant at lag 1. Thus, the optimum ARMA models can be of the order with combinations $p=0, 1$ and $q=0, 1$. Hence, the candidate ARIMA models are of order $(1, 1, 0)$, $(0, 1, 1)$ and $(1, 1, 1)$ for both rice and wheat. These candidate models were fit for the data and AIC and BIC values were found out (Table 3). Based on the AIC and BIC values, ARIMA $(0, 1, 1)$ and ARIMA $(1, 1, 0)$

were selected as the best models for rice and wheat respectively, since the AIC and BIC values were least for these models. These models are given below:

$$\Delta y_{rt} = 1.54 + I\varepsilon_{t-1} \quad (0.056) \quad (0.302) \quad (5)$$

$$\Delta y_{wt} = 1.80 + 0.436y_{wt-1} \quad (0.358) \quad (0.140) \quad (6)$$

where, Δy_{rt} is the first differenced production value for rice at time t and Δy_{wt} is the first differenced production value for wheat at time t . The values in the parenthesis are the standard errors of the corresponding parameter estimates. Model in equation (5) pertains to rice wherein the values for rice production can be obtained by using first lag of moving average terms. Similarly, equation (6) pertains to wheat wherein the values for wheat production can be obtained by using first lag of auto-regression terms.

Table 3. Information Coefficients for ARIMA models

Model	Rice		Wheat	
	AIC	BIC	AIC	BIC
ARIMA (1, 1, 1)	290.01	295.50	248.96	254.45
ARIMA (1, 1, 0)	296.25	299.90	247.08	250.74
ARIMA (0, 1, 1)	288.20	291.85	247.75	251.40

For fitting VARMA models, orders upto p=5 and q=5 are considered for the candidate models. Based on the information criterion and number of significant parameters, VARMA (1, 1, 1) is chosen as the best-suited model. The parameter estimates were obtained by employing least squares estimation. The chosen model is given below:

$$\Delta y_{rt} = 1.252 - 0.233\Delta y_{rt-1} + 0.356\Delta y_{wt-1} + 0.788\varepsilon_{rt-1} + 0.042\varepsilon_{wt-1} \quad (7)$$

(0.056) (0.302) (0.371) (0.337) (0.380)

$$\Delta y_{wt} = 1.252 - 0.233\Delta y_{rt-1} + 0.356\Delta y_{wt-1} + 0.788\varepsilon_{rt-1} + 0.042\varepsilon_{wt-1} \quad (8)$$

(0.056) (0.302) (0.371) (0.337) (0.380)

The residuals of the fitted model are checked for presence of autocorrelation and ARCH effects. Table 4 describes how well each equation of the VARMA (1, 1, 1) model fits the data. The Durbin-Watson statistics is close to 2.0 for rice and wheat indicating the absence of autocorrelation between the residuals. Similarly, the insignificant F-values point out the absence of ARCH effects.

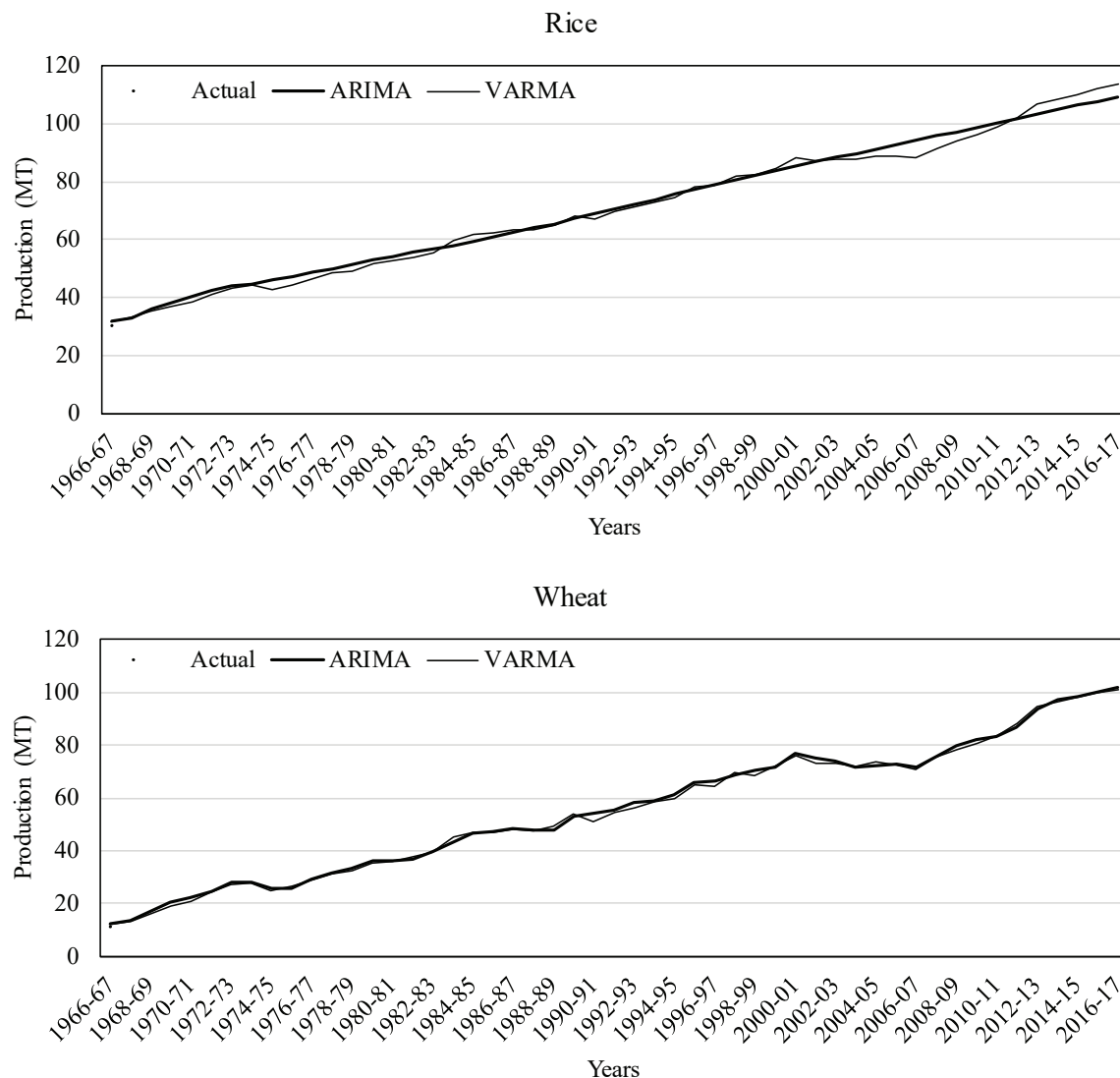


Fig. 4. Actual and forecasted annual rice and wheat production.

Table 4. White noise diagnostics

Variable	Autocorrelation (Durbin-Watson statistics)	ARCH effect	
		F value	p > F
Rice, y_r	1.995	0.18	0.674
Wheat, y_w	2.182	0.13	0.723

The actual and forecasted values are plotted in Fig. 4 for a visual inspection. It is clear from the figure that the fitted values from both ARIMA and VARMA models are close to the actual values. In case of rice, the VARMA model is found to be more efficient than ARIMA model in capturing the fluctuations. But in case of wheat, both ARIMA and VARMA models perform equally well in capturing the fluctuations. The estimates and forecasts of ARIMA and VARMA model are compared using the measures of goodness of fitness via., RMSE and MAPE for different series (Table 5). The results from this are quite exciting. For the data used for training, the VARMA model has performed better for rice while ARIMA model has performed better for wheat. Whereas in case of testing data, VARMA model has performed better for wheat while ARIMA model has performed better for rice. However, the MAPE values are less (<10 %) for both ARIMA and VARMA models. Thus, it is clear from the results that both ARIMA and VARMA model are equally efficient for modelling the time-series data and the best model among them, for the given data, can only be chosen looking at their forecasting performance. Similar results were obtained for modelling demand in emergency department (Patrick *et al.*, 2015). However, VARMA models have been found to perform better than the ARIMA models when the series are cointegrated (Yashavanth *et al.*, 2017). The fact that the production of rice and wheat are independent of each other could be the reason that the VARMA model has not performed better over the ARIMA models. These modelling techniques can be effectively utilized for forecasting agricultural commodities since the production of agricultural commodities are important

and this will help the policy makers for proper policy planning. Policy on import or export of agricultural commodities has to be decided well in advance of the harvesting season. These time-series models can also be utilized for forecasting of prices of essential agricultural commodities as both the consumers and producers need better pricing. Since most of the data are collected over a period of time, these time-series techniques are essential and appropriate for modelling and subsequent forecasting.

4. CONCLUSIONS AND RECOMMENDATIONS

Agriculture in India is dominated by cultivation of two major cereals viz. rice and wheat, which occupy most of the cropped area. Rice is grown in almost all the ecological and agro-climatic regions irrespective of the altitude. Production of rice and wheat have reached all-time high in 2016-17. Since these two are very important crops, it is imperative to model the all-India timeseries data on rice and wheat using various time-series modelling techniques. ARIMA time-series modelling methodology have been utilized widely for modelling and forecasting of univariate time-series data. The multivariate variant of ARIMA, VARMA modelling and forecasting is advantageous over ARIMA for multivariate time-series data. With this background, an attempt was made in this study to model time-series data on rice and wheat separately and jointly by utilizing ARIMA and VARMA time-series modelling methodologies. Generally, it is observed that the VARMA modelling and forecasting methodology performed better than ARIMA with reference to measures of accuracy such as RMSE, MAPE, AIC, and BIC. It may be noted that VARMA performs better only when there exists relation between the variables. Here, the production of rice does not affect the production of wheat and vice-versa. Hence, the performance of these models is not similar with respect to the commodities utilized. Using the best suitable models found for rice and wheat, forecasts for 2019-20 were carried out

Table 5. Comparison of model performance

Variable	Training				Validation			
	RMSE		MAPE (%)		RMSE		MAPE (%)	
	ARIMA	VARMA	ARIMA	VARMA	ARIMA	VARMA	ARIMA	VARMA
Rice, y_r	5.2	5.0	6.6	6.8	1.9	4.5	1.5	3.7
Wheat, y_w	3.3	3.3	6.0	6.2	6.7	6.3	6.1	5.7

and it is found out to be 114 million tons for rice and 106 million tons for wheat. This methodology can be extended for modelling and forecasting of other agricultural commodities and can also be applied in other areas of agricultural research as these techniques have still not been utilized in agriculture and allied disciplines.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**(6), 716-723.
- Badmus, M.A. and Ariyo, O.S. (2011). Forecasting cultivated areas and production of Maize in Nigeria using ARIMA model. *Asian Journal of Agricultural Sciences*, **3**(3), 171-176
- Box, G.E.P. and Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). *Time-Series Analysis: Forecasting and Control*. Pearson education. India.
- Dickey, D.A. and Fuller, W.A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.*, **74**, 427-431
- Gujarati, D.N., Porter, D.C. and Gunasekar, S. (2009). *Basic Econometrics*. Tata McGraw-Hill, New Delhi.
- Hyndman, R.J. and Koehler, A.B. (2006). Another look at measures of forecast accuracy. *Int. J. Forecast.*, **22**(4), 679-688.
- Iqbal, N., Bakhsh, K., Maqbool, A. and Ahmad, A.S. 2005. Use of the ARIMA model for forecasting wheat area and production in Pakistan. *Journal of Agriculture and Social Sciences*, **1**(2), 120-122
- Jambhulkar, N.N. (2013). Modelling of rice production in Punjab using ARIMA model. *International Journal of Scientific Research*, **2**, 1-2
- Kilian L. 2011. Real-time forecasts of the real price of oil. *Journal of Business and Economic Statistics*, **30**(2), 326-36.
- Lütkepohl, H. (2005). Estimation of VARMA Models. In: *New Introduction to Multiple Time-series Analysis*. Springer, Berlin, Heidelberg.
- Lütkepohl, H. (2006). Forecasting with VARMA models. *Handbook of Economic Forecasting* **1**, 287-325.
- Patrick, A.S., Mai, Q., Sanfilippo, F.M., Preen, D.B., Stewart, L.M. and Fatovich, D.M. (2015). A comparison of multivariate and univariate time-series approaches to modelling and forecasting emergency department demand in Western Australia. *Journal of Biomedical Informatics* **57**, 62-73.
- Phillips, P.C.B. and Perron, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika*, **75**(2), 335-346.
- Raghavender, M. (2010). Forecasting rice production in Andhra Pradesh with ARIMA Model. *International Journal of Agricultural and Statistical Sciences*, **6**, 251-258.
- Rahman N.M.F. (2010). Forecasting of boro rice production in Bangladesh: An ARIMA approach. *Journal of Bangladesh Agriculture University*, **8**(1), 103-112
- Ravichandran, S., Muthuraman P. and Rao, P.R. (2012). Time - Series modelling and forecasting India's rice production - ARIMA Vs STM modelling approaches. *International Journal of Agricultural and Statistical Sciences*, **8**, 305-311.
- Sathianandan, T.V. (2007). Vector time-series modelling of marine fish landings in Kerala. *Journal of the Marine Biological Association of India*, **49**(2), 197-205.
- Suleman, N. and Sarpang, S. (2011). Forecasting milled rice production in Ghana using Box-Jenkins approach. *International Journal of Agricultural Management and Development* **2**(2), 79-84.
- Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461-464.
- Tiao, G. and Tsay, R. (1989). Model specification in multivariate time-series (with discussions), *J. Roy. Statist. Soc., B* **(51)**, 157-213.
- Yashavanth, B.S., Singh, K.N., Paul, A.K. and Paul, R.K. (2017). Forecasting prices of coffee seeds using Vector Autoregressive Time-series Model. *Ind. J. Agril. Sci.*, **87**(6), 754-758.