Indian Journal of Agricultural Sciences **90** (7): 1302–5, July 2020/Article

Forecasting long range dependent time series with exogenous variable using ARFIMAX model

KRISHNA PADA SARKAR[1], K N SINGH[2], AMRIT KUMAR PAUL[3], RAMA SUBRAMANIAN V[4], MUKESH KUMAR[5], ACHAL LAMA[6]* and BISHAL GURUNG[7]

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 09 September 2019; Accepted: 11 October 2019

ABSTRACT

Time series analysis and forecasting is one of the challenging issues of statistical modelling. Modelling of price and forecasting is a vital matter of concern for both the farming community and policy makers, especially in agriculture. Many practical agricultural data, principally commodity price data shows the typical feature of long memory process or long range dependency. For capturing the long memory behavior of the data Autoregressive Fractionally Integrated Moving Average (ARFIMA) model is generally used. Sometimes, in time series data besides the original series, data on some auxiliary or exogenous variables may be available or can be made available with a lower cost; like besides the market prices of commodities, market arrivals for that commodity may be available and it affects the market price of commodities. This type of exogenous variable may be incorporated in existing model to improve the model performance and forecasting accuracy, like Autoregressive Fractionally Integrated Moving Average with exogenous variables (ARFIMAX) model. In the present study undertaken at ICAR-IASRI, New Delhi during 2019, daily maximum and modal price of potato of Agra market of UP, India are taken along with daily market arrival. Both the ARFIMA and ARFIMAX model with market arrival as exogenous variable are applied for the data under study. Comparative studies of the fitted models are employed by using the Relative Mean Absolute Percentage Error (RMAPE) and Root Mean Square Error (RMSE) criteria. We could establish superiority of the ARFIMAX model over the ARFIMA model in terms of modeling and forecasting efficiency.

Key words: ARFIMA, ARFIMAX, Exogenous variable, Forecasting

India is largely an agriculture based country where almost 118.7 million people are belong to farming community (according to Census Report 2011). This population consists of 79% of small and medium farmers and 14% of landless farmers. The farmer's welfare depends completely upon the economy of the agricultural sector. Some of the related factors controlling the economy include the availability and the price of the commodities. In this respect time series forecasting is an important and valuable area. It is important in the sense there are so many problems regarding forecasting that involve a time component. There are some practical instances where current value of a variable depends upon the distant past, is the possible indication of long range dependency. Under this dependency instead of Box-Jenkin's ARIMA model, the ARFIMA model (Granger and Joyeux 1980) is used to capture this type of long range dependency as well as for forecasting. Again the ARFIMA model can be improved by incorporating one more exogenous variables in the existing models.

Potato is a chief vegetable crop in India and in recent times there is a huge fluctuation of its market arrival and price. In case of potato market arrival largely be influenced by on the availability of cold storage facility. Often a good harvest of potato may be exposed lower price. It has been seen that potato prices are high during the September-December and low during the month of January-August. So, it is necessary to give a good forecast of its price so that farmers which constitute nearly 82% of the population of India, can decide when to sale, where to sale, how much to sale for getting better market price.

*Corresponding author e-mail: e mail id: chllm6@gmail.com

MATERIALS AND METHODS

Keeping the objectives of the present investigation potato market price and arrival data are collected from National Horticultural Research and Development Foundation (NHRDF) and analyzed at ICAR-IASRI, New Delhi-110012 during 2019. The ARFIMA and ARFIMAX models are fitted in the data set, a details description of the models is given below. Parameters of the models are estimated by Maximum Likelihood Estimation (MLE) technique.

ARFIMA model

The ARFIMA model is generally used for modeling the long memory time series data, because generally used

ARIMA model cannot model the long memory behavior of the time series. In ARIMA model differencing parameter takes only integer value, but ARFIMA model allows d to take fractional value in the range of $-0.5 \le d \le 0.5$. For $d = 0$, the process is stationary and for $0 < d \le 0.5$, the process is said to be stationary and having long memory. Let y_t; (t = 1, 2 ..., n) is a stationary process with mean μ and variance σ^2. Then the ARFIMA model of order (p, d, q), denoted by ARFIMA (p, d, q) can be represented as (Granger and Joyeux 1980)

$$\varphi(L)\,(1-L)^d y_t = \theta\,(L) e_t$$

where e_t is an i.i.d random variable having zero mean and constant variance σ^2. L is the lag operator, d is the fractional difference operator known as long memory parameter, $\varphi(L)$ and $\theta(L)$ are the finite Autoregressive (AR) and Moving Average (MA) polynomials of order p and q respectively. The model has total p + q + 3 parameters μ, σ^2, d, $\varphi = (\varphi_1, \varphi_2, ..., \varphi_p)$ and $\theta = (\theta_1, \theta_2, ..., \theta_q)'$. The parameters are restricted in R^{p+q+3} dimensional space in such a way that following conditions are satisfied (Durham *et al.* 2019)

i. The roots of $\varphi(L)$ and $\theta(L)$ are strictly outside the unit circle.

ii. $|d| < \frac{1}{2}$

iii. $\varphi(L)$ has no repeated roots and $\varphi_p \neq 0$

iv. $\theta^2 > 0$

The parameter of this model is estimated by maximum likelihood estimation (MLE) technique. The exact likelihood function based on n observations $y_n = (y_1, y_2, ..., y_n)'$ is given by:

$$f(y_n / \eta = (2\pi\sigma^2)^{-n/2} |V_n|^{-1/2} \exp\left[\frac{-(y_n - \mu 1_n)' V_n^{-1} (y_n - \mu 1_n)}{}\right]$$

where, $\eta = (\varphi, \theta, d, \mu, \sigma^2)$ is the vector of dimension (p + q + 3)) and $\sigma^2 V_n$ is the variance covariance matrix of Y_n. ML estimates are obtained by maximizing the above likelihood or log likelihood function.

ARFIMAX model

As already discussed ARFIMA model is suitable for long memory time series and exogenous variable can be incorporated in the time series model for better performance of the model. These exogenous variables cab also be incorporated in ARFIMA model which results in Autoregressive Fractionally Integrated Moving Average Model with Exogenous variable (ARFIMAX). Time series model with exogenous variable was initially introduced by Bierens in 1987 by incorporating exogenous variable in ARMA structure (Bierens, 1987). Following Degiannakis, 2008, the ARFIMAX model with k exogenous variable cab be written as (Degiannakis 2008)

$$\varphi(L)\,(1-L)^d\,(y_t - \mu\, x'_t \beta) = \theta\,(L) e_t$$

where, the notations are same as of ARFIMA model and satisfying the conditions given by Durham *et al.* Here x_t is the vector of k exogenous variables at time *t*, β is the vector of coefficients corresponding to k exogenous variables. The model has total of p + q + k + 3 parameters μ, σ^2, d, $\varphi = (\varphi_1, \varphi_2, ..., \varphi_p)' = (\theta_1, \theta_2, ..., \theta q)'$ and $\beta = (\beta_1, \beta_2, ..., \beta_k)'$.

The likelihood function based on the *n* observations on study variable $y, y_n = (y_1, y_2, ..., y_n)'$ and on each *k* exogenous variable $x_1, x_2, ..., x_k$ is given by:

$$L(y_n / \eta^*) = (2\pi\sigma^2)^{-n/2} |V_n|^{-1/2}$$
$$\exp\left[\frac{-(y_n - \mu 1_n - X\beta)' V_n^{-1} (y_n - \mu 1_n - X\beta)}{2\sigma^2}\right]$$

where, $\eta^* = (\varphi, \theta, \mu, d, \beta, \sigma^2)$ be a vector of $(p + q + k + 3)$ parameters, $\boldsymbol{X}$ is $n \times k$ matrix of covariates, $\beta = (\beta_0, \beta_1, ..., \beta_{k-1})'$ and $\sigma^2 V_n$ is the variance covariance matrix of the sample observations.

Before fitting the models stationarity of the data have to check. Augmented Dickey-Fuller (Dickey and Fuller 1979) test and Phillips-Perron (Phillips and Perron 1988) test are employed for the present study.

ADF test

The ADF test based on the ARMA structure of the data set. In this test hypothesis null hypothesis of non-stationarity is tested against the alternative of stationarity. The test regression model for this test is given by

$$\Delta y_t = \boldsymbol{b}' \boldsymbol{x}_t + \alpha y_{t-1} + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t$$

where, Δ is the differencing operator, x_t is the vector of deterministic terms and p is the lag order. The value of p is chosen in such a way that the error terms ε_t become serially uncorrelated. Under the null hypothesis, the ADF t-statistic become equivalent to usual t-statistic for testing $\alpha = 0$.

PP test

The null and alternative hypothesis under PP test are same as of ADF test. The test regression model for this `test is given by

$$\Delta y_t = b' x_t + \alpha y_{t-1} + \varepsilon_t$$

where, ε_t is an white noise process. Asymptotic distribution of PP test statistic under null hypothesis is same as of ADF test statistic. PP test is generally preferred over ADF test due to its robustness and PP test also works under heteroscedastic error and it does not necessitates the lag length in test regression.

Stationarity is checked for both the data under interest as well as exogenous variable. Then the long memory property of the data is checked using ACF and well known Sperio test (Reisen 1994). The analysis has been carried out in R 3.5 software.

RESULTS AND DISCUSSION

For the present study the daily maximum and modal price (₹/q) data of potato of Agra market of Uttar Pradesh,

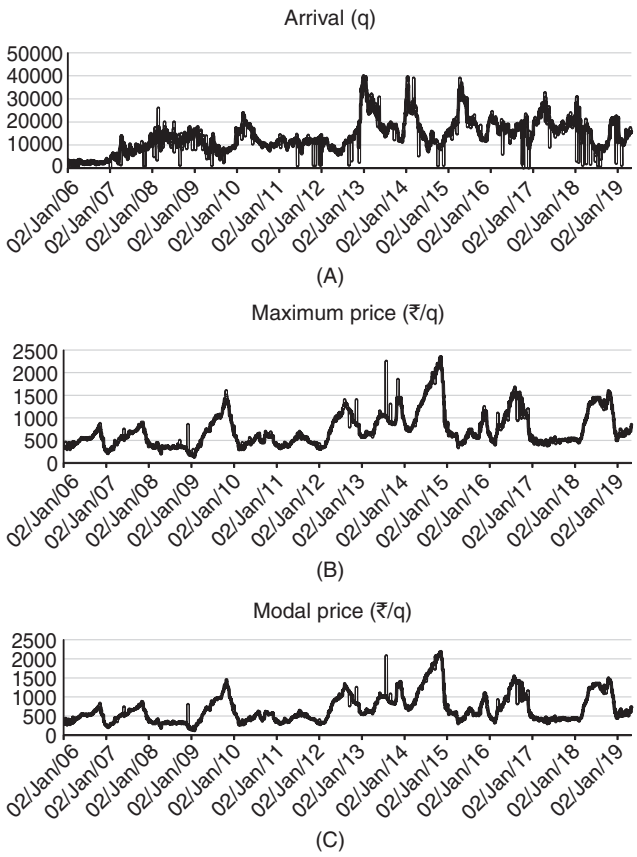

Fig 1 Time plot of potato market arrival (A), maximum (B)and modal (C) price series.

India for the period of January 2006 to June, 2019 are taken from National Horticultural Research and Development Foundation (http://nhrdf.org/en-us/). The daily market arrival (q) has also been taken for the same time period as exogenous variable. The data series consists of 3486 data points from where last 60 observations are used for model validation purpose. Time plots for the series under study are plotted and shown in Fig 1. For checking the stationarity ADF and PP test have been employed and it has found that all the series are stationary. The correlation between maximum price and arrival data are calculated and it is found to be -0.04 (p value 0.02) and that for modal price is -0.05 (p value 0.001) indicating that both the correlations are statistically significant.

The long range dependency of the data series is visualized by plotting ACF of the data series and it has found that autocorrelations up to 120 lags are significant. Long memory parameter is tested by Sperio test and the estimated value of long memory parameter d is 0.41 and 0.46 for maximum and modal price series respectively, indicating presence of persistence long memory in both the data series.

After confirming the presence of long memory in the data set ARFIMA model ARFIMAX model with arrival as exogenous variable is fitted using MLE technique. The parameter estimates, their standard errors and significant p value for the best fitted ARFIMA and ARFIMAX models are given in Table 1 and Table 2 respectively .The AIC and BIC values for the models are calculated for checking the well-fitting of model and it has found that ARFIMAX model better fitted than the ARFIMA model.

Table 1 Parameter estimates of ARFIMA model

Parameter	Estimate	Std. error	p-Value
Maximum Series (2,d,1)			
Constant	730.15	331.96	0.027
AR 1	0.839	0.027	<0.001
AR 2	0.146	0.025	<0.001
d	0.349	0.048	<0.001
MA 1	0.701	0.034	<0.001
Modal Series (1,d,1)			
Constant	682.59	251.05	0.006
AR 1	0.991	0.003	<0.001
d	0.248	0.003	<0.001
MA 1	0.657	0.029	<0.001

Table 2 Parameter estimates of ARFIMAX model

Parameter	Estimate	Std. error	p Value
Maximum price series (2,d,1)			
Constant	759.54	331.81	0.022
AR 1	0.834	0.027	<0.001
AR 2	0.150	0.025	<0.001
d	0.346	0.047	<0.001
MA 1	0.705	0.033	<0.001
Arrival	-0.0031	0.0004	<0.001
Modal price series(2,d,1)			
Constant	722.49	397.11	0.068
AR 1	0.825	0.027	<0.001
AR 2	0.158	0.025	<0.001
d	0.381	0.051	<0.001
MA 1	0.696	0.036	<0.001
Arrival	-0.0028	0.0004	<0.001

For validation and comparison of ARFIMA with ARFIMAX, Root Mean Absolute Percentage Error (RMAPE) and Root Mean Squared Error (RMSE) values are calculated for the last 20 observations and given in Table 3. The results show that the ARFIMAX model perform better over ARFIMA model for both the data set.

For checking the predictive accuracy of the fitted models Diebold-Mariano (DM) test is performed and the test results indicate the superiority of ARFIMAX model over ARFIMA model in terms of predictive accuracy. Residuals for all the models are tested using Ljung Box Test (1978) and it has found that there is no serial autocorrelation for all the models.

From fitting ARFIMA and ARFIMAX models for the data set under study, it has seen that most of the parameters are statistically significant. The exogenous variable market arrival is also significant in ARFIMAX models indicating

Table 3 RMAPE and RMSE values for the fitted models

Data series	RMAPE (%)		RMSE	
	ARFIMA	ARFIMAX	ARFIMA	ARFIMAX
Maximum series	21.10	20.70	217.18	212.58
Modal series	20.80	20.10	186.55	178.13

positive indication of incorporation of this variable in the model. From AIC and BIC values it has found that ARFIMAX model better fitted than the ARFIMA model for both the data set. The RMAPE and RMSE values are also less in ARFIMAX model than ARFIMA model indicating enhanced forecasting accuracy of ARFIMAX model than ARFIMA model. DM test results also suggest that ARFIMAX model has better predictive accuracy over ARFIMA model. The Ljung Box test confirmed the absence of autocorrelation in the residuals, which indicates the proper specification of the model for forecasting of potato price. This study can be extended to other datasets possessing long memory property.

REFERENCES

Assaf A. 2006. Persistence and long-range dependence in the emerging stock market of Kuait. *Middle East Business and Economic Review* **18**(1): 1–17.

Beran J. 1995. *Statistics for Long-Memory Processes*. Chapman and Hall Publishing Inc., New York.

Bierens H J. 1987. ARMAX model specification testing, with an application to unemployment in the Netherlands. *Journal of Econometrics* **35**(1): 161–190.

Box G E P, Jenkins G M and Reinsel G C. 2007. *Time-Series Analysis: Forecasting and Control,* 4th edition. Willey Publication.

Brockwell P J and Davis R A. 1991. *Time Series: Theory and Methods*, 2nd Edition. Springer, New York.

Cheong C W, Isa Z and Nor A H S M. 2007. Modelling financial observable-volatility using long memory models. *Applied Financial Economics Letters* **3**(3): 201–208.

Contreras-Reyes J E and Palma W. 2013. Statistical analysis of autoregressive fractionally integrated moving average models in R. *Computational Statistics* **28**(5): 2309–2331.

Cools M, Moons E, and Wets G. 2009. Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models: assessing the effect of holidays on two site locations. *Transportation Research Record* **2136**(1): 57–66.

Degiannakis S. 2008. ARFIMAX and ARFIMAX-TARCH realized volatility modeling. *Journal of Applied Statistics* **35**(10): 1169–1180.

Dickey D and Fuller W. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of American Statistical Association* **74**: 427–431.

Durham G, Geweke J, Porter-Hudak S and Sowell F. 2019. Bayesian Inference for ARFIMA Models. *Journal of Time Series Analysis* **40(4):** 388–410

Granger C W J and Joyeux R. 1980. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* **4**: 221–238.

Hao Y, Wu J, Sun Q, Zhu Y, Liu Y, Li Z, and Yeh T C J. 2013. Simulating effect of anthropogenic activities and climate variation on Liulin Springs discharge depletion by using the ARIMAX model. *Hydrological Processes* **27**(18): 2605–2613.

Hosking J R M. 1981. Fractional differencing. *Biometrika* **68**: 165–176.

Ljung G M and Box G E P. 1978. On a measure of a lack of fit in time series models. *Biometrika* **65**(2): 297–303.

Paul R K, Gurung B and Paul A K. 2014. Modelling and forecasting of retail price of arhar dal in Karnal, Haryana. *Indian Journal of Agricultural Sciences* **85**(1): 69–72.

Paul R K, Gurung B and Samanta S. 2015a. Monte Carlo simulation for comparison of different estimators of long memory parameter: An application of ARFIMA model for forecasting commodity price. *Model Assisted Statistics and Applications* **10**(2): 117–128.

Paul R K, Prajneshu and Ghosh H. 2014. Development of out-of-sample forecast formulae for ARIMAX-GARCH model and their application. *Journal of Indian Society of Agricultural Statistics* **68**(1): 85–92.

Paul R K, Prajneshu and Ghosh H.2013. Modelling and forecasting of wheat yield data based on weather variables. *Indian Journal of Agricultural Sciences* **83**: 180–183.

Paul R K. 2014. Forecasting wholesale price of pigeonpea using long memory time-series models. *Agricultural Economics Research Review* **27**(2): 167–176.

Phillips P C B and Perron P. 1988. Testing for unit roots in time series regression. *Biometrika* **75**: 335–346.

Reisen V A. 1994. Estimation of the fractional difference parameter in the ARIMA (*p*, *d*, *q*) model using the smoothed periodogram. *Journal of Time Series Analysis* **15**(3): 335–350.

Sriboonchitta S, Chaiboonsri C, Chaitip P, Chaitip A, Kovacs S, and Balogh P. 2013. An investigation on the international tourists' expenditures in Thailand: A Modelling Approach. *Applied Studies in Agribusiness and Commerce* **7**(1)**:** 15–18.