FORECASTING BUFFALO MILK PRODUCTION IN INDIA: TIME SERIES APPROACH

Yashavanth Basavapatna Subbanna*, Sanjiv Kumar and Sharath Kumar Maddur Puttaraju

ABSTRACT

Globally, India stands first both in production as well as consumption of milk. Nearly 50% of the Indian milk production comes from buffalo followed by cow and goat. India is home to buffalo with approximately 56% of world buffalo population. Given the importance of buffalo milk, an attempt was made to model and forecast the annual buffalo milk production using various time series analysis techniques. The Autoregressive Integrated Moving Average (ARIMA) model, the Artificial Neural Network (ANN) model and ARIMA-ANN hybrid model were used to model the time series data of 58 years collected from secondary sources. Among the three models, the ARIMA-ANN hybrid model was found to be the best for the data under consideration based on forecast accuracy measures. This is because of the ability of ARIMA-ANN hybrid model to capture both linear and nonlinear structures in the data. By using the ARIMA-ANN hybrid model, the annual buffalo milk production was forecasted and it was found to exceed 1,000 million tonnes in the coming years.

Keywords: *Bubalus bubalis*, buffaloes, ARIMA, forecasting, hybrid model, neural network

INTRODUCTION

India ranks first in milk production and consumption in the world with milch animal population of 125.34 million. The milk production in India has increased from 55.6 million tonnes in 1991 to 1992 to 187.7 million tonnes in 2018 to 2019, reporting 237.58 % growth. During the last decade, milk production in India has recorded a growth of about 4.8% CAGR (Mishra *et al*., 2019). Similarly, the per capita availability of milk also has increased from 130 gram/day during 1950 to 1951 to 374 gram/day in 2017 to 2018, which is more than the world estimated average consumption during 2017. The increase in demand has also necessitated the increase in milk production (Kumar *et al*., 2014). India reported 21.29% share of the total milk production in the world during 2017. As per the 2019 livestock census, India has 192.49 million cattle and 109.85 million buffaloes population. Nearly half of the milk production in India comes from buffaloes (35% from indigenous buffaloes and 14% from non-descript buffaloes) (DAHD Annual Report, 2018-2019). It is estimated that half of the milk produced is converted into traditional milk product in India. Among milk from different species, the buffalo milk is considered to yield better quantity and quality of traditional dairy products (Minhas *et al*., 2002; Zicarelli, 2004;

Indian Council of Agricultural Research, National Academy of Agricultural Research Management, Hyderabad, India, *E-mail: yashavanthbs@gmail.com

Kumar *et al*., 2010).

Dairy being the largest contributor to Agriculture GDP, forecasting of milk production is required to know the availability and need of milk so that necessary policy intervention can be made to meet this gap (Mishra *et al*., 2019). Efforts have already been made to study and forecast the milk production in India (Paul *et al*., 2014; Mishra *et al*., 2019), Pakistan (Ahmed *et al*., 2011), cow milk production in Ethiopia (Taye *et al*., 2020) and buffalo milk production in Brazil (Saude *et al*., 2020). However, given the significant contribution of buffalo milk to the overall milk production in India, there have been no efforts to study the production trends of buffalo milk in India. Moreover, these studies have been carried out using only Autoregressive Integrated Moving Average (ARIMA) methodology of time series analysis which is efficient only for linear time series data. With this background, an attempt is made to study and forecast the production of buffalo milk using proven time series modelling techniques such as ARIMA, Artificial Neural Network Models (ANN) and the ARIMA-ANN hybrid model which can capture linear component, nonlinear component and both linear and nonlinear components, respectively. Various studies (Ravichandran *et al*., 2018; Ray *et al*., 2016) have established the supremacy of hybrid model in forecasting time series data related to agriculture and allied sectors.

MATERIALS AND METHODS

Data

The time series data used for forecasting comprises the annual buffalo milk production in India for the period of 1960 to 1961 to 2018 to 2019. The data was obtained from FAOSTAT. For the analysis, out of the total available 58 years' data, the first 50 years' data (1960 to 1961 to 2010 to 2011, 85% of available data) are used for training the models and the last 8 years' data (2011 to 2012 to 2018 2019, 15% of available data) are used for model validation.

Time series analysis techniques

There are various time series models in literature which are to be used based on the characteristics of the data. The most popular time series model used for short term forecasting are the ARIMA models which are an extension to the ARMA models. In an ARIMA model, it is assumed that the time series under study is a linear function of the past values and random shocks. An ARIMA model, represented as ARIMA (p,d,q), comprises three components: *p*, the order of Auto-Regression (AR); *d*, the order of integration (or differencing) to achieve stationarity; and *q*, the order of Moving Average (MA). An ARMA (p, q) process is defined by Equation

$$y_t = \mu+\varphi_1 y_{t-1}+\varphi_2 y_{t-2}+\cdots+\varphi_p y_{t-p}+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q}+\varepsilon_t$$

where, y_t are the actual value and ε_t are random shocks at time period t. μ, φ_i (i=1, 2,…, *p*) and θ_i (j=1, 2,…,*q*) are the model parameters. The random errors, ε_t are assumed to be independently and identically distributed with a mean of zero and a constant variance of σ^2.

The first and most important requirement in ARIMA modeling is to ensure that the series under study is stationary since the estimation procedure is available only for a stationary series. A series is considered as stationary if its mean and the autocorrelation structures do not change over time. The stationarity of a time series can be

confirmed either by a time plot or by using unit root tests. If a series is found to be non-stationary based on these tests, it can be made stationary either by differencing or by transformations. The number of times a series is differenced to achieve stationarity is referred to as the order of integration/ differencing, *d*. After achieving the stationarity of the data series, the 4 step Box-Jenkins approach *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 Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), respectively. This is followed by the estimation stage, where parameters of the identified tentative models (candidate 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) or Bayesian Information Criterion (BIC) 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.

Despite being widely used, ARIMA models are best suited only for linear time series data as they fail to capture the non-linear structures. A wide variety of ANN models are used in such cases where non-linear structures are to be captured. The popular Multi-Layer Perceptron (MLP) networks with two layers, one hidden and one output layer connected acyclically are very often used for non-linear time series modelling. In the MLP networks, the relation between the output y_t and inputs y_{t-1}, y_{t-2} ,…, y_{t-p} is as below:

$$y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j . g(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij}\, y_{t-1}) + \varepsilon_t$$

where α and β are the model parameters, *p* is the number of input nodes, *q* is the number of hidden nodes and *g* is the transfer function. In case of ANN autoregressive models (ANNAR), the lagged variables y_{t-i} (i=1,2,…,*p*) are the inputs. Such an ANNAR model is represented as NNAR (p, q). Among several transfer functions, the logistic function is most often used.

According to Zhang (2003), a time series is composed of a linear autocorrelation structure and nonlinear component as $y_t = L_t + N_t$ where L_t and N_t denote the linear and non-linear components, respectively. Hence, a hybrid model which can capture both linear and non-linear components may perform better than the individual models. Therefore, we attempted to use the ARIMA-ANN hybrid model for fitting and forecasting annual buffalo milk production. Building a hybrid ARIMA-ANN model consists of two steps. In the first step, the linear component L_t is modelled by using ARIMA. In the second step, the residuals of the ARIMA model containing information on the non-linearity of the series are modelled through an ANN. Subsequently, the best suited ARIMA and ANN models are combined to obtain the hybrid model.

Forecast evaluation measures

The ability of different models to forecast the time series values is assessed by using forecast evaluation measures. The two common measures

are the root mean squared error (RMSE) and the mean absolute percentage error (MAPE). The RMSE measures the overall performance of a model and is given

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

where, y_t is the actual value for time $\hat{y}_t$, is the predicted value for time t, and n is the number of predictions. The second criterion, the mean absolute percentage error is a measure of average error for each point forecast and is given by

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

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.

RESULTS AND DISCUSSION

The foremost step in time series analysis is to plot the data under study to have a visual inspection over its behaviour. Figure 1 shows the time series plot of annual production of buffalo milk production in India during the period 1960-1961 to 2018-2019.

A perusal of Figure 1 indicates an increasing trend in the annual buffalo milk production. The Mann-Kendal test for trend also confirms the presence of positive trend ($z = 10.854$, $P<0.01$). The Figure 1 also gives the CAGR in each year for buffalo milk production. It is observed that the highest CAGR (12.41%) was reported in 1983 whereas the least CAGR (-7.99) was reported in the year 1967. In the past two decades, the CAGR is seen to be fluctuating between 2 to 6 %. The Augmented Dickey-Fuller (ADF) test is performed to check the stationarity of the data, results of which are given the Table 1. The Table also includes the results of ADF test performed after first and second differencing. The values clearly indicate the non-stationarity of original series as well as after first differencing at 1% level of significance. The series was found to be stationary after second differencing. The time plots of the differenced series are presented in Figure 2 from which it is evident that the first order differenced series is showing some positive trend whereas the second order differenced series is stationary.

After obtaining the stationary series, the candidate ARIMA models were identified based on the Autocorrelation and Partial Autocorrelation functions (ACF and PACF). The ACF and PACF plots (Figure 3) indicate that the maximum order for AR is 1 and for MA is 2, since autocorrelation and partial autocorrelation coefficients are not significant for higher orders. Accordingly, 2 candidate ARIMA models *viz*, ARIMA (1,2,1) and ARIMA (1,2,2) are identified. Among these two candidate models, ARIMA (1,2,1) is chosen as the appropriate model based on the Information Criteria. The candidate models and their respective AIC and SBC values are given in Table 2. The Ljung-Box test was carried out to check the adequacy of the ARIMA (1,2,1) model and it was observed that the model is adequate ($Q^* = 14.641$, $P = 0.06$). The residual series, ACF and histogram of the residual series obtained from ARIMA (1,2,1) model are presented in Figure 3. It is pertinent from the figure that the residual series is normally distributed with no significant autocorrelation.

Subsequently, the series was modelled using Neural Networks. The NNAR (1,1) model was found to be the best among all other higher order

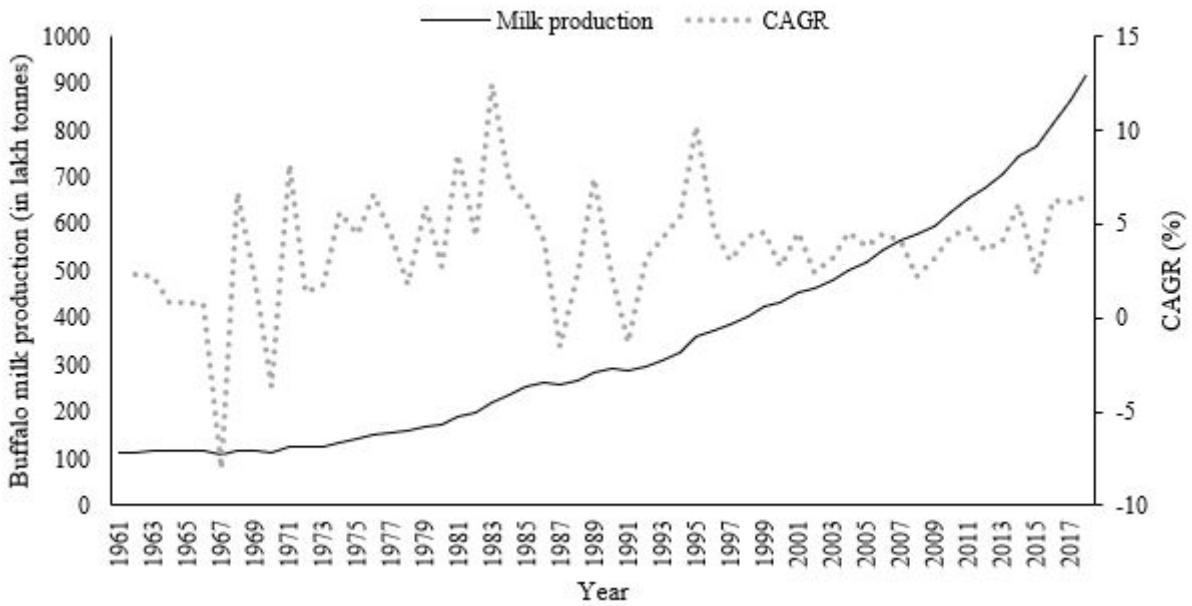

Figure 1. Time plot of the annual buffalo milk production and CAGR.

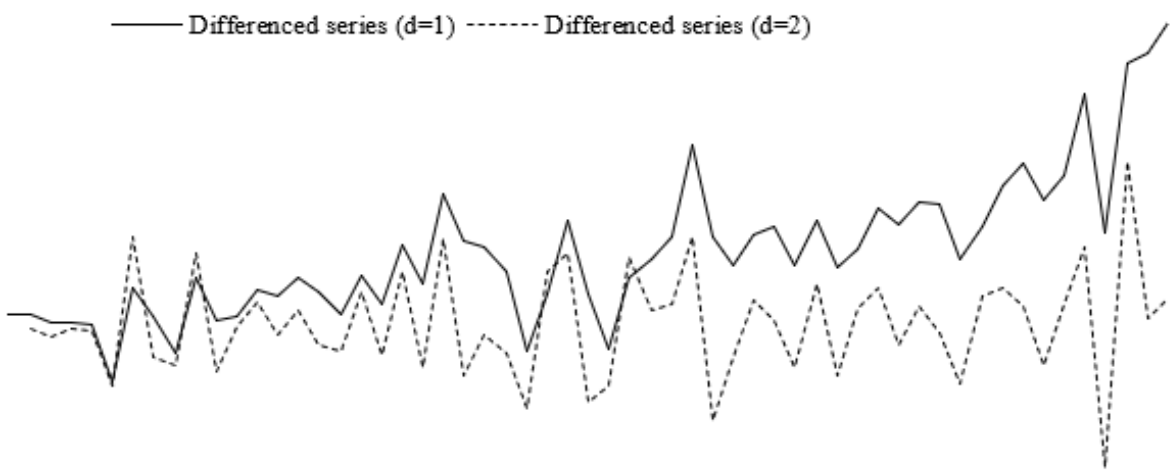

Figure 2. Time plot of the differenced series of annual buffalo milk production.

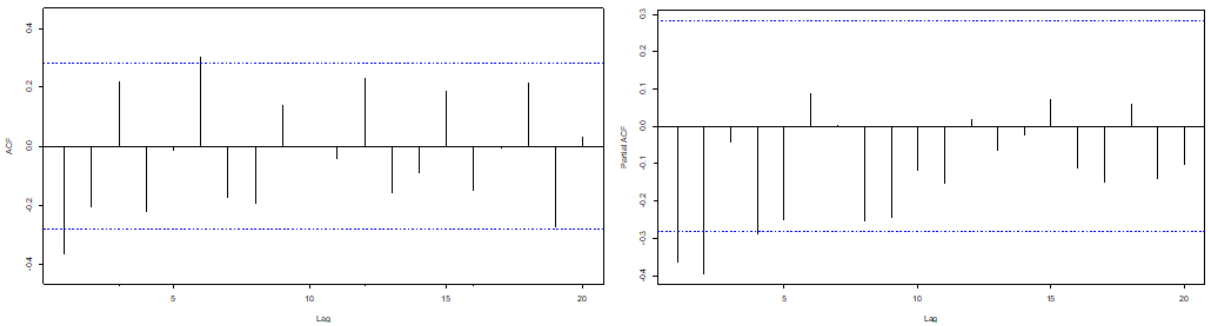

Figure 3. Autocorrelation and partial autocorrelation plots.

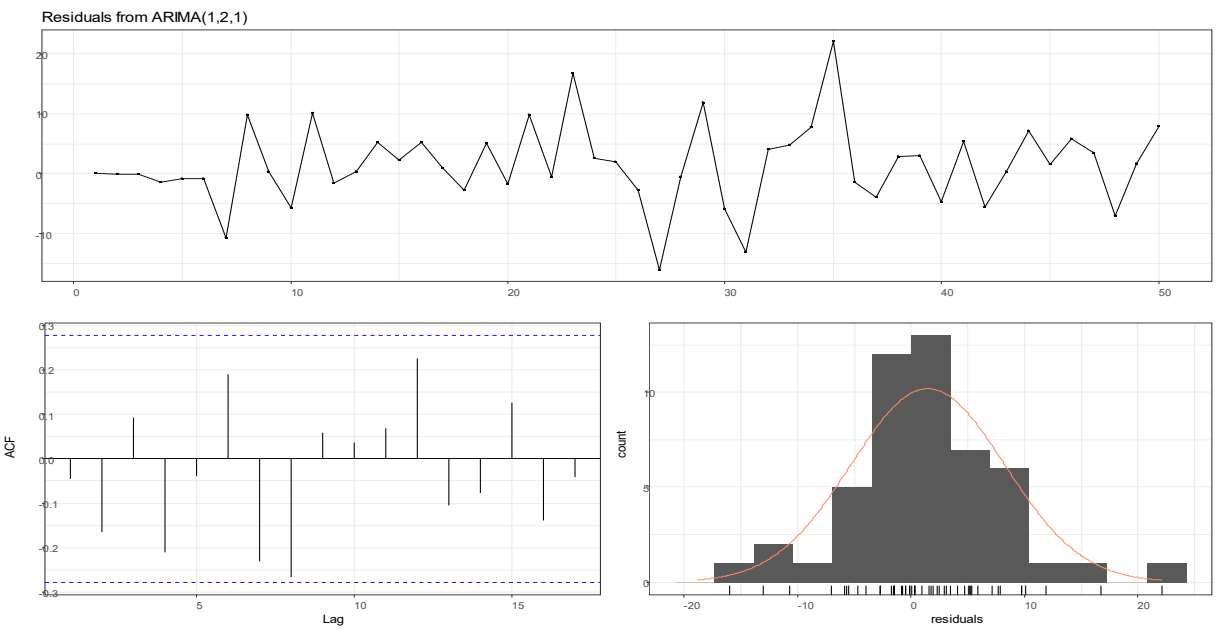

Figure 4. Time plot of residual series, its ACF plot and histogram.

Table 1. Results of the augmented Dickey-Fuller test for stationarity.

Series	ADF statistic	P-value
Original	-1.082	0.916
1st order differenced	-4.120	0.012
2nd order differenced	-5.291	<0.001

Table 2. Candidate ARIMA models, their parameters and information coefficients.

Candidate model	Model equation	AIC	BIC
ARIMA (1,2,1)	$\hat{y}_t'' = 0.132y''_{t-1} - 0.79\varepsilon_{(t-1)}$	329.94	335.56
ARIMA (1,2,2)	$\hat{y}_t'' = - 0.516y''_{t-1} - 0.082\varepsilon_{(t-1)} - 0.546\varepsilon_{(t-2)}$	330.49	337.98

'' indicates that the series are differenced twice to achieve stationarity

Table 3. Measures of goodness of fit for different models.

Model	Training data		Testing data	
	MAPE	RMSE	MAPE	RMSE
ARIMA	2.164	6.868	6.501	068.008
NNAR	2.257	6.751	12.146	121.478
ARIMA-NN Hybrid	2.192	6.642	6.266	66.600

Table 4. Forecasted annual buffalo milk production values for 2019-2020 to 2023-2024.

Year	Forecasted values		
	Point forecast	95 % lower limit	95% upper limit
2019-2020	967.344	950.544	984.143
2020-2021	1,015.619	986.536	1,044.702
2021-2022	1,063.878	1,021.271	1,106.484
2022-2023	1,112.138	1,054.704	1,169.571
2023-2024	1,160.397	1,086.882	1,233.912

Neural Network models. The ARIMA models are known to efficiently capture the linear structures whereas the neural network models are efficient in capturing the non-linear structures in the series. To capture both linear and non-linear structures, the ARIMA-ANN hybrid models are tried out in which the residuals obtained from the ARIMA model are once again modelled using the neural network model. The NNAR (1,1) model was found suitable for modelling the residual series obtained from the ARIMA (1,2,1) model. Later, both ARIMA (1,2,1) model for the data series and the NNAR (1,1) model for the residual series are combined to obtain the hybrid model. The performance of all the three models, viz., ARIMA, ANN and ARIMA-ANN hybrid models is evaluated using forecast evaluation measures, namely, RMSE and MAPE (Table 3). The results indicated that the ARIMA-ANN hybrid model is best suited for modelling and forecasting the annual buffalo milk production since it was successful in capturing both the linear and non-linear structures. Using the most efficient model *viz*, hybrid ARIMA-ANN model, the annual buffalo milk production was forecasted for the next 5 years (Table 4). For forecasting, the complete study data set (1960 to 1961 to 2018 to 2019) was used. From the forecasted values, it is observed that the annual buffalo milk production is going to exceed 1,000 lakh tonnes in the year 2020 to 2021.

CONCLUSION

Time series analysis techniques are being used extensively used for analyzing data from agricultural and allied sectors including animal sciences. These techniques have been found to be useful in foreseeing the repercussions that may happen in the future and thus help in being prepared to face the situation. Given the substantial contribution of buffalo milk to India's overall milk production, this study attempts to model and forecast the annual buffalo milk production using time series analysis techniques. The ARIMA model which is extensively used for such studies is capable of capturing only linear component in the data. Hence, in this study, the nonlinear ANN technique as well as the ARIMA-ANN hybrid technique which are capable of capturing nonlinear and both linear and nonlinear components, respectively, have been used in addition to ARIMA model. Among all the three models, the ARIMA(1,2,1)-NNAR(1,1) hybrid model was found to be efficient in modelling the annual buffalo milk production as per the forecast evaluation measures, *viz*, MAPE and RMSE. By using the ARIMA(1,2,1)-NNAR(1,1) hybrid model, the annual buffalo milk production was forecasted for the next 5 years and it is estimated that the annual buffalo milk production will surpass 1,000 lakh tonnes in the year 2020 to 2021.

REFERENCES

Ahmed, F., H. Shah, I. Raza and A. Saboor. 2011. Forecasting milk production in Pakistan. *Pak. J. Agric. Res*., **24**(1-4): 82-85. Available on: http://pjar.org.pk/Issues/Vol24_2011No1_4/Vol24No1_4Page82.pdf

Box, G.E.P., G.M. Jenkins and G.C. Reinsel. 1994. *Time Series Analysis, Forecasting and Control*, 3rd ed. Prentice Hall, Englewood Clifs, New Jersey, USA.

DAHD. 2019. *Department of Animal Husbandry and Dairying Annual Report*. Available on: https://dahd.nic.in/sites/default/filess/Annual%20Report.pdf

Kumar, A., P.K. Joshi, P. Kumar and S. Parappurathu. 2014. Trends in the consumption of milk and milk products in India: implications for self-sufficiency in milk production. *Food Secur.*, **6**(5): 719-726. DOI: 10.1007/s12571-014-0376-y

Kumar, R. and R. Singh. 2010. Buffalo production system in India. *Revista Veterinaria*, **21**(Sppl. 1): 1-13.

Minhas, K.S., J.S. Sidhu, G.S. Mudahar and A.K. Singh. 2002. Flow behavior characteristics of ice cream mix made with buffalo milk and various stabilizers. *Plant Food. Hum. Nutr.*, **57**(1): 25-40. DOI: 10.1023/a:1013106116587

Mishra, P., C. Fatih, H.K. Niranjan, S. Tiwari, M. Devi and A. Dubey. 2019. Modelling and forecasting of milk production in Chhattisgarh and India. *Indian J. Anim. Res.*, **54**: 912-917. DOI: 10.18805/ijar.B-3918

Paul, R.K., W. Alam and A.K. Paul. 2014. Prospects of livestock and dairy production in India under time series framework. *Indian J. Anim. Sci.*, **84**(4): 462-466.

Ravichandran, S., B.S. Yashavanth and K. Kareemulla. 2018. Oilseeds production and yield forecasting using ARIMA-ANN modelling. *Journal of Oilseeds Research*, **35**(1): 57-62.

Ray, M., A. Rai, V. Ramasubramanian and K.N. Singh. 2016. ARIMA-WNN hybrid model for forecasting wheat yield time-series data. *Journal of the Indian Society of Agricultural Statistics*, **70**(1): 63-70. Available on: http://isas.org.in/jsp/volume/vol70/issue1/8%20-%20Mrinmoy.pdf

Saude, L.M.S., G.T. Gabriel and P.P. Balestrassi. 2020. Forecasting of buffalo milk in a Brazilian diary using the ARIMA model. *Buffalo Bull.*, **39**(2): 201-213. Available on: https://kuojs.lib.ku.ac.th/index.php/BufBu/article/view/550

Taye, B.A., A.A. Alene, A.K. Nega and B.G. Yirsaw. 2020. Time series analysis of cow milk production at Andassa dairy farm, West Gojam zone, Amhara region, Ethiopia. *Modeling Earth Systems and Environment*, 7: 181-189. DOI: 10.1007/s40808-020-00946-z

Zicarelli, L. 2004. Buffalo milk: Its properties, dairy yield and mozzarella production. *Vet. Res. Commun.*, **28**(Sppl. 1): 127-135. DOI: 10.1023/B:VERC.0000045390.81982.4d

Zhang, G. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, **50**: 159-175. DOI: 10.1016/S0925-2312(01)00702-0