DR-1782
[1-6]

RESEARCH ARTICLE Asian Journal of Dairy and Food Research, Volume Issue : ()

ARIMA-Genetic Algorithm Approach for Forecasting Milk Production in India

Pramit Pandit[1], Bishvajit Bakshi[2], Moumita Paul[1], B.S. Pooja[3]

10.18805/ajdfr.DR-1782

ABSTRACT

Background: Dairying in India has witnessed a radical transformation from a largely unorganised activity into a thriving organised industry. However, there are only a limited number of earlier attempts that specifically evaluated the milk production pattern of India. Moreover, most of these earlier studies are quite dated and have employed autoregressive integrated moving average (ARIMA) models, which suffer from the problem of local optima. To overcome this lacuna, we have utilised a genetic algorithm (GA) in the ARIMA framework. Its suitability to forecast the annual milk production in India has also been assessed comparatively with respect to the traditional ARIMA methodology.

Methods: For the current study, data on annual milk production (in million tonnes) in India for the period from 1980 to 2019 have been utilised. For both the approaches under investigation, the whole data series is first divided into two sets, namely the training set and the testing set. The production data of 1980-2016 have been utilised for the model building purpose while retaining the last 3 years' data for the post-sample evaluation.

Result: Outcomes emanated from the post-sample assessment clearly suggest that the ARIMA-GA approach has outperformed the traditional ARIMA methodology in terms of root mean square error (RMSE) and mean absolute percentage error (MAPE) values. It is also evident that GA has substantially minimised the error related to the parameter estimation.

Key words: ARIMA, Evolutionary algorithms, Forecasting, GA, Milk production.

INTRODUCTION

An ever-expanding population, sustainable economic development, urbanisation and upsurging health consciousness are accelerating the shift in food consumption patterns in India. Over the last few years, the per capita consumption of food grains has either remained almost constant or revealed a declining trend, whereas the consumption of high-value livestock food commodities has shown a sharp increase (Kumar *et al.*, 2007). Milk and milk products have been amongst the most important components of the Indian food basket, with their share of monthly per capita food expenditure increasing from 11.50 to 14.90 per cent in rural areas and from 15.70 to 18.40 percent in urban areas between 1983 and 2010 (GoI, 2010). Even though dairying in India has witnessed a radical transformation from a largely unorganised activity into a thriving organised industry, the majority of the agricultural research has been concentrated on food grain production. Consequently, a thorough understanding of the milk production pattern of India is necessitated not only for academic explorations but also for policy formulations and interventions (Kumar *et al.*, 2014).

However, there are only a limited number of earlier attempts that specifically evaluated the milk production pattern of India. Pal *et al.* (2007) have made a comparative evaluation between double exponential smoothing technique and autoregressive integrated moving average (ARIMA) model for forecasting milk production in India. Their study has indicated the superiority of the ARIMA model over the other one. Paul *et al.* (2014) have employed ARIMA (1, 1, 0) model for modelling and forecasting milk production in India. Deshmukh and Paramasivam (2016) have also forecasted the same with ARIMA and Vector autoregression (VAR) models and concluded that ARIMA (1, 1, 1) model is more suitable in terms of lesser root mean square error (RMSE) and mean absolute percentage error (MAPE) values. Apparently, the literature review suggests that the ARIMA model has been dominating time series analysis in this sector. Notwithstanding its sheer power and dominance, the traditional ARIMA methodology suffers from the problem of local optima (Ong *et al.*, 2005). To overcome this lacuna, Ervural *et al.* (2016) have recently introduced a genetic

[1]Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur-741 252, West Bengal, India.
[2]Centre for Management of Health Services, Indian Institute of Management, Ahmedabad-380 015, Gujarat, India.
[3]Department of Data Science, Prasanna School of Public Health, Manipal Academy of Higher Education, Manipal-576 104, Karnataka, India.

Corresponding Author: Bishvajit Bakshi, Centre for Management of Health Services, Indian Institute of Management, Ahmedabad-380 015, Gujarat, India.
Email: bishvajitb93@gmail.com

How to cite this article: Pandit, P., Bakshi, B., Paul, M. and Pooja, B.S. (2022). ARIMA-Genetic Algorithm Approach for Forecasting Milk Production in India. Asian Journal of Dairy and Food Research. DOI: 10.18805/ajdfr.DR-1782.

Submitted: 07-07-2021 **Accepted:** 04-03-2022 **Online:** 09-04-2022

algorithm (GA)-based estimation of autoregressive moving average (ARMA) model in the field of natural gas consumption forecasting. The above facts clearly indicate that there is a lack of systematic investigation on forecasting milk production in India in light of the recent advancements. Moreover, the earlier studies are also quite dated. With this backdrop, an attempt has been made in this study to assess the suitability of the ARIMA-GA approach for forecasting annual milk production in India in comparison to the traditional ARIMA methodology.

MATERIALS AND METHODS

Data

For the current study on annual milk production (in million tonnes) in India, the data series for the period from 1980 to 2019 has been collected and compiled from the various issues of 'Basic Animal Husbandry Statistics' published by the Department of Animal Husbandry and Dairying, Ministry of Fisheries, Animal Husbandry and Dairying, Government of India, New Delhi and from the website of Ministry of Agriculture and Farmers Welfare, Government of India. Methodologically, the whole data series is first divided into two sets, namely the training set and the testing set. The production data of 1980-2016 have been utilised for the model building purpose while retaining the last 3 years' data for the post-sample evaluation.

ARIMA model

The time series variable in an ARMA model is considered to be a linear function of its past values and random shocks. It includes both autoregressive and moving average processes to obtain greater flexibility in the fitting of actual time series data. An ARMA (p, q) model can be specified as (Box *et al.*, 2015):

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

Where,

y_t and e_t are the actual observation and random error at time t, respectively; ϕ_i (i=1,2,...,p) and θ_j (j=1,2,...,q) are the model parameters. p and q, being integers, are referred to as the order of the model. Random shocks ε_t are assumed to be independent and identically distributed with zero mean and a constant variance σ^2.

A popular generalisation of ARMA models, which incorporates a wide class of non-stationary time series models, is achieved by introducing the concept of differencing into the model. An ARIMA model representing homogeneous non-stationary behaviour can be written as follows:

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right)(1 - B)^d y_t = \left(1 - \sum_{j=1}^{q} \theta_j B^j\right)\varepsilon_t$$

Where,

B is the backshift operator defined as $By_t = y_{t-1}$ and d represents the order of differencing. In practice, d is usually 0, 1, or at most 2. The Box-jenkins methodology (*i.e.*, the ARIMA methodology) includes three iterative steps, namely identification, parameter estimation and diagnostic checking (Durdu, 2010).

Identification

The first step in ARIMA model building is to check and ensure stationarity of the series being analysed as the estimation procedures are available only for stationary series. In the next step, based on autocorrelation and partial autocorrelation patterns, one or several potential models are identified.

Parameter estimation

At the identification stage, one or more tentative models that appear to provide adequate statistical representations of the available data are chosen. Once a tentative model is specified, model parameters are estimated by the method of maximum likelihood estimation (MLE).

Diagnostic checking

In this step, the white noise test for the residuals of the tentatively selected model is carried out. If residuals are not white noise, again a candidate model is selected and the same procedure is repeated until a valid model is found.

ARIMA-genetic algorithm approach

GA, developed by John Holland and his collaborators (Holland, 1992), is an abstraction of biological evolution based on Charles Darwin's theory of natural selection (Darwin, 1964). It is one of the most extensively used evolutionary algorithms in terms of the diversity of its applications; from graph colouring to pattern recognition, from financial markets to multi-objective engineering optimisation problems, *etc.* Holland is widely credited as being the first to utilise crossover, recombination, mutation and selection in the study of adaptive and artificial systems. As a matter of fact, the GA is incomplete as a problem-solving approach without these genetic operators. GAs have multiple edges over the classic optimisation algorithms. Parallelism and the ability to handle complex problems are two of its most remarkable features.

GAs basically rely on the survival of the best individuals. Over generations, the fitness function improves and the best solution is obtained finally. Building GA for the ARIMA problem, as presented in Fig 1, includes a number of steps (Ding, 2011; Ervural *et al.*, 2016; Rathod *et al.*, 2017; Yang, 2020).

String representation

Each chromosome consists of two parts to represent AR (p) and MA (q), with each dimension, equal to the length (p + q).

Initial population

The initial population is chosen at random. The number of chromosomes in each generation is referred to as the population size and it is a key parameter for improving the performance of GA. However, there is no specific standard to determine the size.

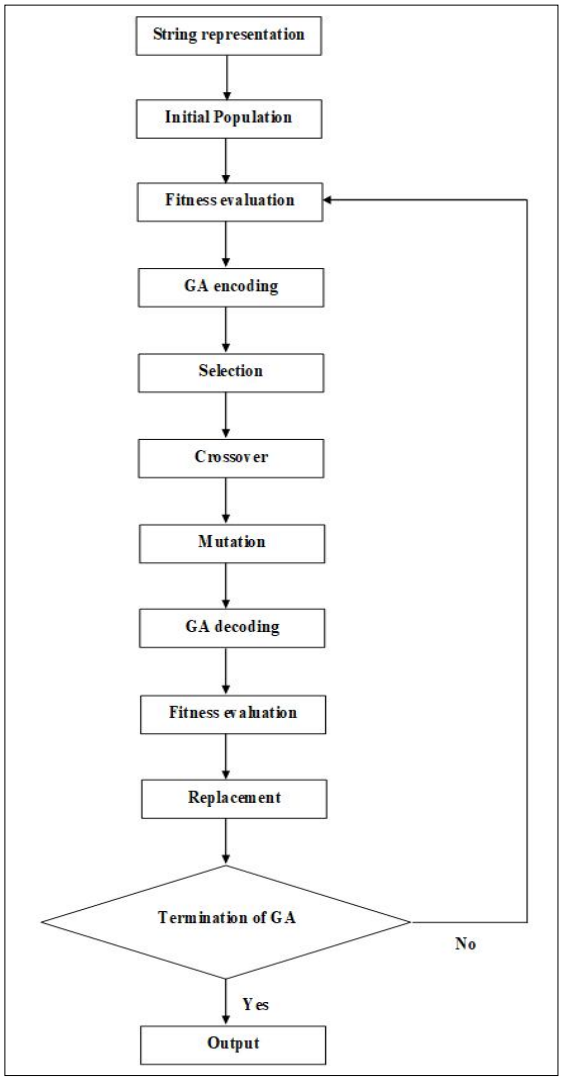

Fig 1: Schematic diagram of genetic algorithm.

Fitness evaluation

To initiate the estimation procedure, an objective function should be defined for evaluation in terms of the fitness function. In the current study, the function is specified in terms of MAPE, such as:

$$\text{Fitness} = \frac{1}{1 + \text{MAPE}}$$

GA encoding

In order to perform crossover and mutation, the real values have to be represented as binary strings (0 and 1). The number of bits (n) in each variable is provided by:

$$2^n = \frac{\text{Range}}{\text{Quality}}$$

Where,

The quality value is often set to 0.001.

Selection

The fitness function is used to select chromosomes from the current population to create new offspring for the next generation. The higher the fitness, the more likely the chromosome contributes one or more offspring to the next generation. Choosing a small number of chromosomes restricts the number of offspring in the next generation while retaining too many chromosomes can lead to undesirable traits in the next generation. Hence, a minimum of 50% is kept in natural selection. Among the several methods available for the selection operation (Haupt and Haupt, 2004), we have used the randomisation-based Roulette wheel selection in this study.

Crossover

Crossover provides new offspring for the next generation by exchanging information between two randomly selected parent chromosomes. Crossover substantially improves GA in terms of exploration and diversification abilities with a view to obtaining the global optimum point. However, in most cases, crossover is not performed on all of the selected chromosomes. In practice, the choice of crossover probability is made between 0.6 and 1.0.

Mutation

After performing crossover, this random search, *i.e.*, the mutation is applied to each offspring in order to avert a premature convergence. It can be depicted as a random bit with a small probability typically ranging between 0.1 and 0.001, which is selected at random from the total number of bits from the population matrix.

GA decoding

To carry out further fitness evaluation, the string values are converted into their equivalent real values by decoding. This process is performed by utilising the equation:

$$X = X_{lower} + \frac{X_{dec}}{2^n - 1}(X_{upper} - X_{lower})$$

Where,

x and X_{dec} represent the real and the decimal decoded value of the gene, respectively. X_{lower} and X_{upper} accordingly indicate the lower and upper bound of x.

Fitness evaluation

Once the selection, crossover, mutation and decoding are performed, evaluation of the new offspring is carried out in the earlier fashion.

Replacement

After evaluation, the parents need to be replaced by the new offspring. The replacement operations can be categorised into two main types, *viz.*, generational/non-overlapping replacement and overlapping replacement. In the former one, the parent population is replaced by the offspring population except for the best individuals in parents, whereas in the case of the latter one, according to their fitness values, both the offspring and parent population compete to survive into the next generation.

Termination of GA

Once the convergence criterion is met, such as the maximum number of iterations is reached or the desired fitness value

is obtained, the GA is terminated. Otherwise, *i.e.*, if not met, the entire algorithm is repeated until the desired fitness value is obtained.

Assessment of forecasting accuracy

The forecasting ability of both models is assessed in terms of the two widely used accuracy measures, *viz.*, RMSE and MAPE (Wong *et al.*, 2005; Gonzalez-Vidal *et al.*, 2019). RMSE measures the overall performance of a model and has the form:

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

Where,

y_t and $\hat{y}_t$ represent the t^{th} actual and predicted value in the test data set, respectively and n denotes the size of the test data set. The second measure, *i.e.*, MAPE is a measure of per cent average error for each point forecast and is given by:

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

Where,

The notations have the same meanings as above.

RESULTS AND DISCUSSION

Summary statistics of the time series data under study are reported in Table 1. Average annual milk production in India is observed to be 85.53 million tonnes during the study period. High CV value indicates the presence of instability in the series. The foremost step in time series analysis is to plot the data. Fig 2 shows the time series plot of annual milk production (in million tonnes) in India from 1980 to 2019. The upward trend over time suggests the non-stationary nature of the series.

To further confirm this claim, we have applied the augmented dickey-fuller (ADF) test (Paparoditis and Politis, 2018). Results of the ADF test, as presented in Table 2, have confirmed the non-stationarity and stationarity of the level and first differenced series, respectively. The plots of autocorrelation function (ACF) and partial autocorrelation function (PACF) of the first differenced series, as provided in Fig 3, are utilised to tentatively identify the order of the ARIMA model. Among the candidate models, ARIMA (1, 1, 1) model has been chosen as appropriate on the basis of minimum Akaike information criterion (AIC) and Bayesian information criterion (BIC) values (Swain, 2018). The parameter estimates of the selected model by MLE are provided in Table 3. Result of the Ljung-Box test (test statistic value = 0.01, p-value = 0.97) and the ACF, PACF plots of the residuals (Fig 4) have validated the well-behaved nature of the ARIMA (1, 1, 1) residuals.

Following the traditional ARIMA model building, an attempt has been made to utilise the GA to estimate the ARIMA parameters. The optimised parameters of GA with respect to minimisation of the objective function have been resulted after several runs. The model summary of the ARIMA-GA approach is provided in Table 4. Table 5 provides the comparative results for the traditional ARIMA and ARIMA-GA approaches in terms of the post-sample RMSE and MAPE values. Outcomes emanated from the post-sample assessment clearly suggest that the ARIMA-GA approach has outperformed the traditional ARIMA methodology as the employment of GA has substantially minimised the error related to the parameter estimation. The observed and the ARIMA-GA predicted milk production in India are graphically presented in Fig 5.

Table 1: Descriptive statistics of annual milk production (in million tonnes) in India.

Statistic	Value
Mean	85.53
Minimum	30.40
Maximum	187.70
Standard deviation	42.84
Coefficient of variation (%)	50.08
Skewness	0.74
Kurtosis	-0.34

Table 2: Results of the ADF test.

Series	Test statistic	p value
Level	-0.26	0.99
First difference	-4.00	0.02

Table 3: Parameter estimates of the ARIMA model by MLE method.

Parameter	Estimate	p value
μ	4.23	0.02
ϕ_1	0.93	<0.01
θ_1	-0.15	<0.01

Table 4: Model summary of ARIMA-GA.

Parameter	Optimum values
Model specifications	
Population size	200
Selection type	Roulette wheel
Selection rate	50%
Crossover type	Single point
Crossover rate	65%
Mutation rate	0.05
Iteration	23
Parameter estimates	
Parameter	Estimate
μ	3.93
ϕ_1	1.04
θ_1	-0.03

Table 5: Post-sample assessment of model accuracy.

Model	RMSE	MAPE
ARIMA	4.69	2.22
ARIMA-GA	0.36	0.19

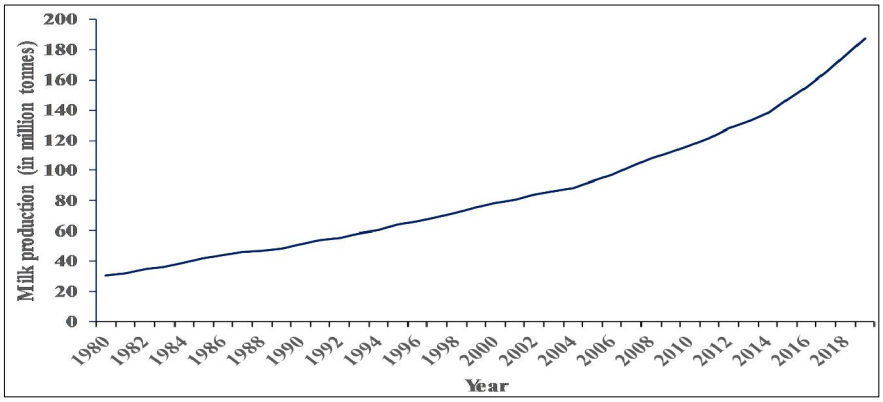

Fig 2: Annual milk production (in million tonnes) in India during the study period.

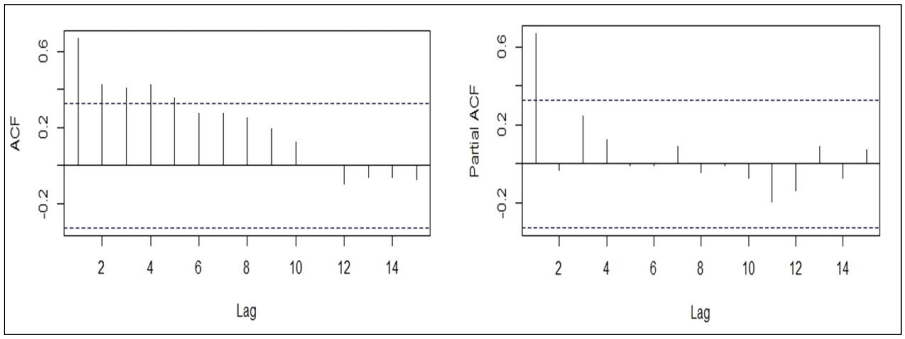

Fig 3: ACF and PACF plots of the first differenced series.

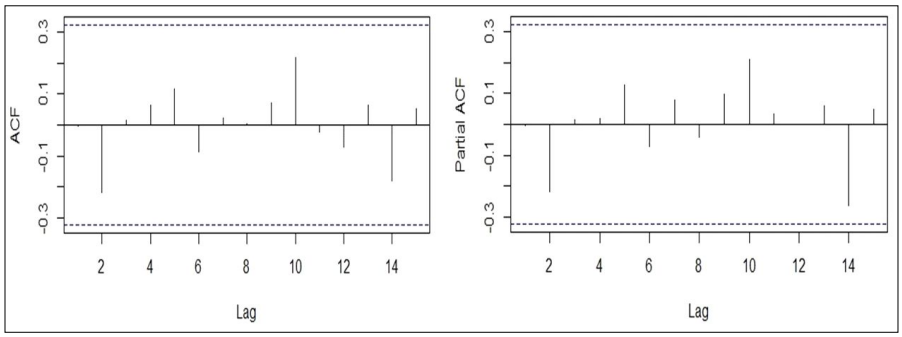

Fig 4: ACF and PACF plots of the residual series.

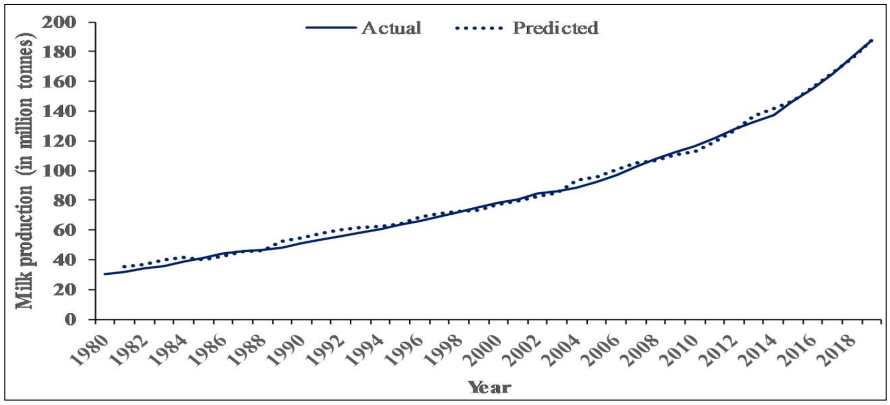

Fig 5: Observed and the ARIMA-GA predicted milk production in India.

CONCLUSION

In order to overcome the drawbacks of traditional ARIMA models that are generally employed for forecasting milk production, the genetic algorithm is gaining momentum. The study has compared the traditional ARIMA methodology with the ARIMA-GA approach for forecasting milk production in India. The outcomes reveal that the ARIMA-GA approach has performed substantially better than the traditional one. Consequently, it can be also inferred that forecasting using GA has been proven to be more accurate. However, the generalisation of the study includes investigations of the performance of other evolutionary algorithms under the ARIMA framework.

Conflict of interest: None.

REFERENCES

Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015). Time series analysis: Forecasting and control. John Wiley and Sons.

Darwin, C. (1964). On the Origin of Species: A Facsimile of the First Edition. Harvard University Press.

Deshmukh, S.S. and Paramasivam, R. (2016). Forecasting of milk production in India with ARIMA and VAR time series models. Asian Journal of Dairy and Food Research. 35(1): 17-22.

Ding, S., Su, C. and Yu, J. (2011). An optimizing BP neural network algorithm based on genetic algorithm. Artificial Intelligence Review. 36(2): 153-162.

Durdu, Ö.F. (2010). Application of linear stochastic models for drought forecasting in the Büyük Menderes river basin, western Turkey. Stochastic Environmental Research and Risk Assessment. 24(8): 1145-1162.

Ervural, B.C., Beyca, O.F. and Zaim, S. (2016). Model estimation of ARMA using genetic algorithms: A case study of forecasting natural gas consumption. Procedia-Social and Behavioral Sciences. 235: 537-545.

Gonzalez-Vidal, A., Jimenez, F. and Gomez-Skarmeta, A.F. (2019). A methodology for energy multivariate time series forecasting in smart buildings based on feature selection. Energy and Buildings. 196: 71-82.

Government of India. (2010). 66 round national sample survey of consumption expenditure. National Sample Survey Organization. Government of India, New Delhi.

Haupt, R.L. and Haupt, S.E. (2004). Practical Genetic Algorithms. John Wiley and Sons.

Holland, J.H. (1992). Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. MIT press.

Kumar, A., Joshi, P.K., Kumar, P. and Parappurathu, S. (2014). Trends in the consumption of milk and milk products in India: Implications for self-sufficiency in milk production. Food Security. 6(5): 719-726.

Kumar, P., Dey, M. and Madan, M. (2007). Long-term changes in food basket and nutrition in India. Economic and Political Weekly. 42(35): 3567-3572.

Ong, C.S., Huang, J.J. and Tzeng, G.H. (2005). Model identification of ARIMA family using genetic algorithms. Applied Mathematics and Computation. 164(3): 885-912.

Pal, S., Ramasubramanian, V. and Mehta, S.C. (2007). Statistical models for forecasting milk production in India. Journal of Indian Society of Agricultural Statistics. 61(2): 80-83.

Paparoditis, E. and Politis, D.N. (2018). The asymptotic size and power of the augmented Dickey-Fuller test for a unit root. Econometric Reviews. 37(9): 955-973.

Paul, R.K., Alam, W. and Paul, A.K. (2014). Prospects of livestock and dairy production in India under time series framework. Indian Journal of Animal Sciences. 84(4): 130-134.

Rathod, S., Singh, K.N., Arya, P., Ray, M., Mukherjee, A., Sinha, K., Kumar, P. and Shekhawat, R.S. (2017). Forecasting maize yield using ARIMA-Genetic Algorithm approach. Outlook on Agriculture. 46(4): 265-271.

Swain, S., Nandi, S. and Patel, P. (2018). Development of An ARIMA Model for Monthly Rainfall Forecasting over Khordha District, Odisha, India. In: Recent Findings in Intelligent Computing Techniques. [Sa, P.K., Bakshi, S., Hatzilygeroudis, I.K. and Sahoo, M.N. (eds.)]. pp. 325-331. Springer.

Wong, J.M., Chan, A.P. and Chiang, Y.H. (2005). Time series forecasts of the construction labour market in Hong Kong: The Box Jenkins approach. Construction Management and Economics. 23(9): 979-991.

Yang, X.S. (2020). Nature-inspired Optimization Algorithms. Academic Press.