Article

Outlook on Agriculture
2017, Vol. 46(4) 265–271
© The Author(s) 2017
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0030727017744933
journals.sagepub.com/home/oag

Forecasting maize yield using ARIMA-Genetic Algorithm approach

Santosha Rathod[1], KN Singh[1], Prawin Arya[1], Mrinmoy Ray[1], Anirban Mukherjee[2], Kanchan Sinha[1], Prakash Kumar[1], and Ravindra Singh Shekhawat[1]

Abstract

Maize is widely cultivated throughout the world and has highest production among all the cereals. India is the sixth largest producer of maize in the world, contributing 2% of global production and accounting for 9% of the total food grain production in the country. Based on increasing growth rates of poultry, livestock, fish, and milling industries, the demand for maize is expected to increase from the current level of 17 to 45 million tons by 2030. To understand the growing pattern and economics of crop production, it is necessary to predict crop yield using statistical models and geographic information system soil mapping and the impacts of insect and pest damage. In this study, the focus was to forecast maize yield in India using an autoregressive integrated moving average (ARIMA) model and genetic algorithm (GA) approach. GA simulates the evolution of living organisms, where the fittest individual dominates the weaker ones by mimicking the biological mechanism of evolution, such as selection, crossover, and mutation. GA has successfully been applied to solve optimization problems. The study reveals that implementation of GA in ARIMA enhances the prediction accuracy of the model.

Keywords

Evolution, statistical models, fitness function, India

Introduction

Maize (*Zea mays*) is considered to be the queen of cereal crops and one of the most staple food crops in the world, second only to rice and wheat and is used in animal feed and many industrial applications. It is cultivated widely throughout the world and has highest production among all the cereals. The global production of maize was more than 960 Million Metric tonne (MN MT) in 2013–2014. The crop has tremendous genetic variability, which enables it to thrive in tropical, subtropical, and temperate climates. Global production of maize has grown at a compound annual growth rate (CAGR) of 3.4% over the last decade, from 717 MN MT in 2004–2005 to 967 MN MT in 2013–2014. The area under maize cultivation over the period has increased at a CAGR of 2.2%, from 146 MN ha in 2004–2005 to 177 MN ha in 2013–2014. Productivity of maize has increased at a CAGR of 1.2% from 4.9 MT/ha in 2004–2005 to 5.5 MT/ha in 2013–2014. The United States is the largest maize producer, contributing 37% of global production followed by China (22%). India is the sixth largest producer of maize (2%). In India, maize is the third most important crop after rice and wheat which accounts for 9% of the total food grain production; 85% of the maize crop is cultivated in the kharif season.

Maize production in India has grown at a CAGR of 5.5% over the last decade from 14 MN MT in 2004–2005 to 23 MN MT in 2013–2014. The area under maize cultivation has increased at a CAGR of 2.5% from 7.5 MN ha in 2004–2005 to 9.4 MN ha in 2013–2014. Factors such as its adaptability to diverse agroclimatic conditions, lower labor costs, and lowering of the water table in the rice belt of India have contributed to the increase in the maize cropped area. The productivity of maize (yield) has increased at a CAGR of 2.9% from 1.9 MT/ha in 2004–2005 to 2.5 MT/ha in 2013–2014 (Agricultural situation in India, 2015).

Maize is grown throughout the year in India across a wide range of environments, extending from extreme semi-arid to subhumid and humid regions. The crop is also very popular in the low and mid-hill areas of the western and northeastern regions. In broad terms, maize cultivation can be classified into two production environments: (i)

[1] Division of Forecasting and Agricultural Systems Modeling, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India
[2] Social Science Section, ICAR-Vivekananda Institute of Hill Agricultural Research, Almora, Uttarakhand, India

Corresponding author:
Santosha Rathod, Division of Forecasting and Agricultural Systems Modeling, ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012, India.
Email: santosha.rathod@icar.gov.in

traditional maize growing areas, including Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh and (ii) nontraditional maize areas, viz., Karnataka and Andhra Pradesh. Maize production is dominated by Andhra Pradesh and Karnataka, producing approximately 38% of India's maize in 2010–2011.

In the past, maize was mainly confined to food consumption but changing Indian dietary patterns now mean that it is being largely grown for feed purposes (60%). Based on the increasing growth rate of poultry, livestock, fish, and milling industries, maize demand is expected to increase from the current level of 16.72 to 45 million tons by 2030 (Report of India maize summit, 2015). Constraints on the low productivity of maize include climatic conditions resulting in drought/excess water associated with increased pressure of diseases or pests, imbalance or inefficient use of nutrients, limited adoption of improved production–protection technologies, and deficiencies in the production and distribution system of quality seed. To overcome these challenges, forecasting crop yield using statistical models, predicting the possibilities of insect or pest incidence based on weather forecasting, geographic information system–based soil fertility mapping, and remote sensing for crop productivity or pest incidence assessment are all becoming essential elements to support improved agricultural management. Forecasting is used to provide an aid to decision-making and to improve future planning. Governments are increasingly concerned with accurate crop production forecasts as they provide some idea of the size of the national income, the overseas balance of payments situation, and any marketing difficulties likely to be associated with the sale of products in domestic as well as overseas markets. Statistical forecasting models are often used to develop an appropriate forecast methodology using historical data to predict future trends in growth with the help of identifying patterns within data (Choi et al., 2015).

In this study, effort has been made to forecast the yield of maize in India using an autoregressive integrated moving average (ARIMA) model combined with a genetic algorithm (GA) approach. ARIMA models (Box et al., 1994) have been widely used for crop yield and other agricultural production forecasting. Sarika et al., 2011 conducted a study on modeling and forecasting time series data of pigeon pea production in India using Box–Jenkins ARIMA time series methodology; the ARIMA (2, 1, 0) model was reported to perform better among other models of the ARIMA family for modeling as well as for forecasting purposes. Jambhulkar (2013) applied the ARIMA methodology based on the lowest Akaike information criterion (AIC) and Bayesian informa-tion criterion (BIC) values; ARIMA (1, 1, 2) model found to be superior to other ARIMA models for forecasting rice production in Punjab, India. Tahir and Habib (2013) applied linear trend, quadratic trend, exponential trend, and S-curve models for trend analysis of the area and production of maize in Pakistan and reported that forecast values were very close to observed values with a positive increasing trend in Pakistan. Karim et al. (2010) conducted a study on the growth pattern of wheat production in Bangladesh and showed that the best fitting was quadratic, linear, and cubic models. The analysis found that if the present growth rates continued, then wheat production in Bangladesh would reach 1.54 million tons. Naveena et al. (2014) used ARIMA model for forecasting coconut production in India. The best model ARIMA (1, 1, 1) was selected based on the minimum root mean square error values.

Sometimes in ARIMA, modeling the large sum of the squared residuals in the Ljung–Box statistic indicates that the model is not appropriate for the data. The common zeros in the AR and MA process indicate parameter redundancy, which means that the model can be shortened by two parameters (Rolf et al., 1997). Under such conditions, the ordinary least square (OLS) method cannot converge to the true parameter. To overcome this problem, a powerful iterative optimization technique is required to estimate the parameters of the ARIMA model under consideration. GA is a powerful global searching method developed by Holland (1975) which was successfully applied to solve optimization problems in many areas (e.g. Parviz et al., 2010; Rolf et al., 1997, Zaer et al., 2012). In this study, effort has been made to employ the GA for parameter estimation of the ARIMA model to overcome the shortcomings of the OLS method for forecasting maize yield in India.

ARIMA model building

Box and Jenkins (1970) introduced the concept of ARIMA in 1970 in their book *Time Series Analysis Forecasting and Control*. The technique is used to forecast future values of a series based completely on past values. Here "AR" means lags of the differenced series appearing in the forecasting equation; "MA" is the lag of the forecast errors and a time series which needs to be differenced for making it stationary is termed "integrated." Generally, a nonseasonal ARIMA model, denoted as ARIMA (*p, d, q*), is expressed as

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \ldots + \emptyset_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \tag{1}$$

where Y_t and ε_t are the actual values and random error at time *t*, respectively; $\emptyset_i (i = 1,2,\ldots,p)$ and θ_i (i= $1,2,\ldots,q$) are AR and MA parameters, respectively; and *p* and *q* are integers and often referred to as orders of AR and MA polynomials, respectively. Random errors ε_t are assumed to be independently and identically distributed with mean zero and the constant variance $\sigma_\varepsilon{}^2$. Box and Jenkins (1970) propose a practical three-stage procedure for finding a good model, viz., (i) identification of the model, (ii) parameter estimation, and (iii) diagnostic checking of the model.

Identification. The autocorrelation function (ACF) and partial ACF (PACF) are used to identify the number of potential AR and MA orders to be selected. For testing of stationarity, the most popular methods used include the augmented Dickey–Fuller (ADF) unit root test and the Phillips–Perron unit root tests with a null hypothesis that the time series is not stationary and the alternative hypothesis is that the time series is stationary. Details of these tests are found in the literature (Dickey and Fuller, 1979; Phillips and Perron, 1988).

Estimation. For the estimation phase, the parameters identified in the identification stage are estimated for the ARIMA model by employing an iterative least squares method. AIC and BIC values are used for choosing the best model and are given as follows

$$\mathrm{AIC} = T\log(\sigma^2) + 2(p + q + 1) \tag{2}$$

and

$$\mathrm{BIC} = T\log(\sigma^2) + 2(p + q + 1)\log T \tag{3}$$

where T represents the number of observations utilized for estimation of parameters and σ^2 represents the mean square error.

Diagnostic checking. Based on the ACF and PACF of the residuals, the independency of the residuals can be diagnosed. If the residuals approximate to white noise (residuals of the models are found to be random in nature), the sample space–time ACFs should be effectively zero. The Ljung–Box test (Box et al. 1994) can be employed on the original series or to the residuals after fitting a model with the null hypothesis that the series is white noise, and the alternative hypothesis is that one or more autocorrelations up to lag m are not zero. The test statistic is given by

$$Q^* = T(T + 2)\sum_{k=1}^{m} \frac{{r_k}^2}{T - k} \tag{4}$$

where T is the number of observations deployed to estimate the model and m is the maximum number of lags. The statistics Q^* approximately follows a χ^2 distribution with $(T - k)$ degrees of freedom, where k is the number of parameters estimated in the ARIMA model and r_k is the ACF of residual at lag k. If it is not adequate, we return to the identification stage to speculatively select another model.

GA. A GA (Holland, 1975) is a random search algorithm based on the basic principles of biological evolution and natural selection. The GA simulates the evolution of living organisms, where the fittest individual dominates over the weaker ones, by mimicking the biological mechanism of evolution, such as selection, crossover, and mutation; therefore, it is also known as an evolutionary algorithm. Thus, GA is a powerful optimization technique applicable in many science and engineering problems such as system identification, controller design, neural networks, fuzzy systems, image and signal processing, and motion planning of robot manipulators (e.g. Parviz et al., 2010; Zaer et al., 2012). In GA, the population of possible solutions is evaluated to estimate the best solution based on three main concepts, (i) reproduction, (ii) evaluation, and (iii) selection. The genetic reproduction is performed by two basic genetic operators: crossover and mutation. The evaluation is performed by means of the fitness function that depends on the specific optimization problem. The selection is the process of choosing the best parent individuals according to their relative fitness. Building a GA for any problem includes initialization, fitness evaluation, GA encoding, selection, crossover, mutation, GA decoding, fitness evaluation, and finally termination (Figure 1).

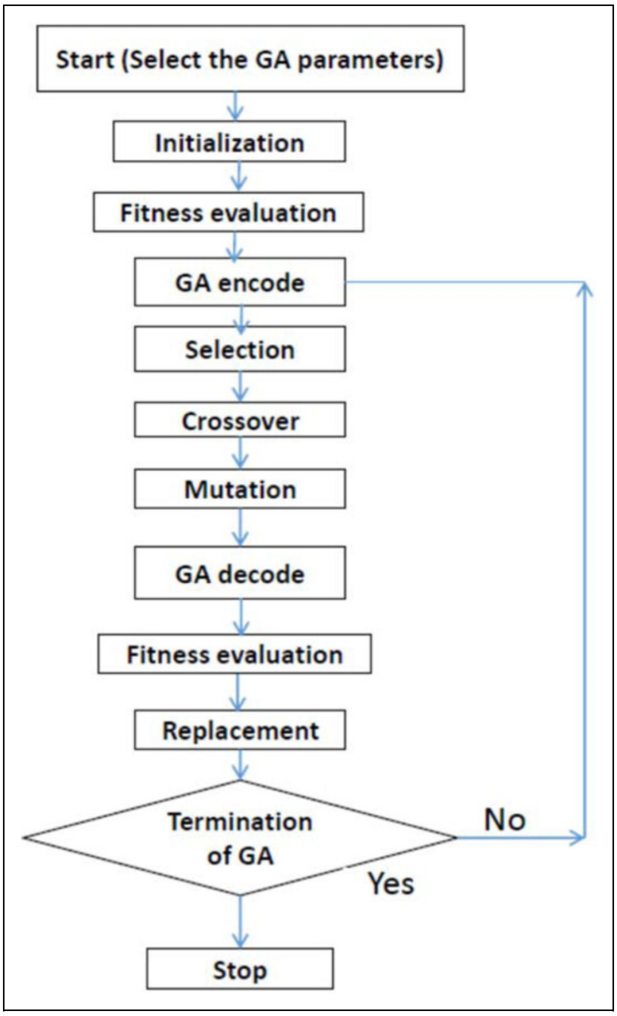

Figure 1. Steps in the genetic algorithm formulation.

Table 1. Summary statistics of maize yield series.

Statistic	Series	Statistic	Series
Observation	64	Skewness	0.76
Mean	1366.30	Kurtosis	–0.29
Standard deviation	529.17	Coefficient of variation (%)	38.73

Steps in GA formulation

Initialization. Based on the pre-assumed number of parameters; the number of AR and MA parameters (p max + q max) are considered as initial

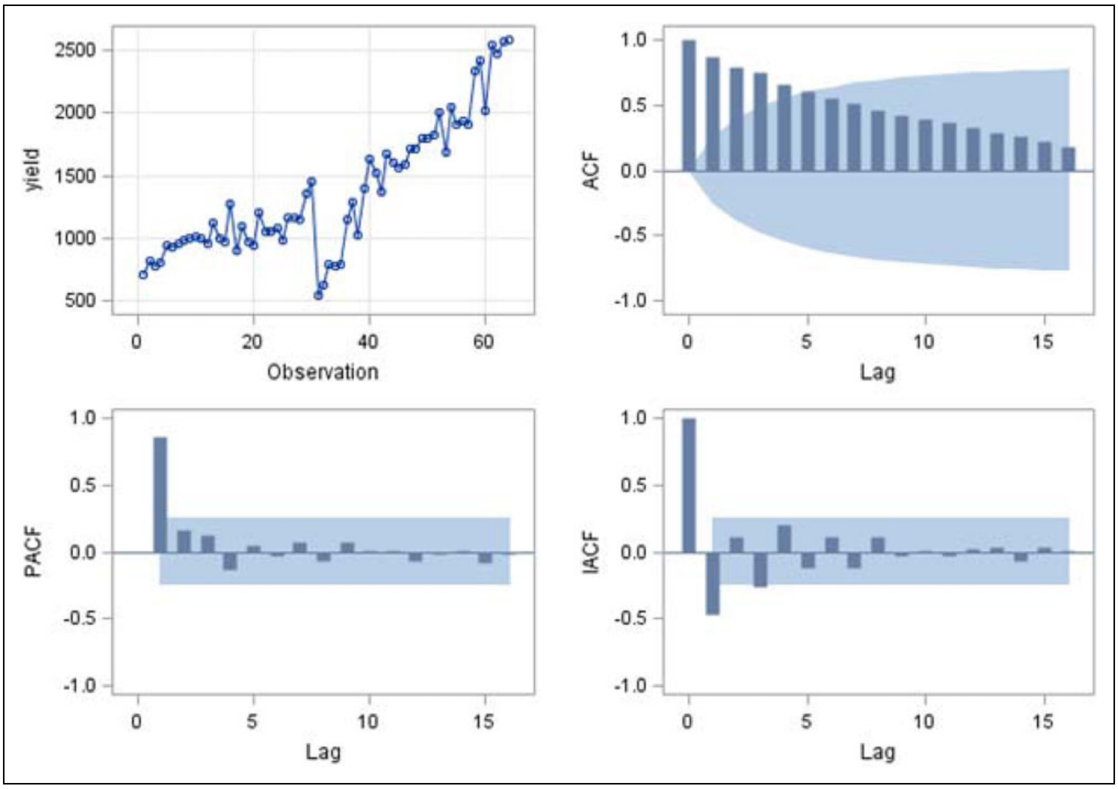

Figure 2. Trend and correlation analyses of actual series.

population, where each parameter is an individual in the population, which consists of different sets of solutions and each solution set is termed a chromosome. Each value in the solution set consists of a number of bits, that is, genes. Consider a matrix of population size (Np), resulting in a population matrix containing $Np \times (p\,\max + q\,\max)$ elements. The roots of the numerator and the denominator must be inside the unit circle to satisfy the stability and invertibility conditions. Thus, the genes will generate within the chosen range and the first generation is randomly generated in real values.

Fitness evaluation. To start with the estimation process, an objective function should be defined in terms of fitness function for evaluation. In general, the fitness function may be a mathematical or experimental function that achieves the desired output. In the proposed method, the function is based on minimizing the difference between the actual and estimated values. In our case, the fitness function was defined in terms of mean absolute percentage error (MAPE) % as

$$\text{Fitness} = \frac{1}{1 + \text{MAPE}} \qquad (5)$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \qquad (6)$$

where Y_t is the actual series and $\hat{Y}_t$ is the estimated value.

GA encoding. To carry out the GA operations, viz., crossover and mutation, the real values must be represented in binary strings 0 and 1. The number of bits (n) in each variable is given by the formula (7) the chromosome variables and the operation is per formed with a certain quality value, that is, 0.001.

Table 2. Testing for stationary in the actual series.

ADF test statistic				PP test statistic			
		Probability				Probability	
Single mean	With trend	Single mean	With trend	Single mean	With trend	Single mean	With trend
−1.05	−7.07	0.98	0.63	−1.03	−17.77	0.88	0.08

ADF: augmented Dickey–Fuller.

$$2^n = \frac{\text{Range}}{\text{Quality}} = \frac{x_{\text{upper}} - x_{\text{lower}}}{\text{Quality}} \qquad (7)$$

Selection. For creating new offspring for the next generation, chromosomes are chosen from the current population based on the fitness function. The larger the fitness, the higher the probability that the chromosome will contribute one or more offspring to the next generation. Choosing few chromosomes limits the availability of offspring in the next generation and keeping too many chromosomes may contribute to undesired traits in the next generation. Therefore, we keep a minimum of 50% in the natural selection. There are several methods for the selection operation (Haupt and Haupt, 2004). In

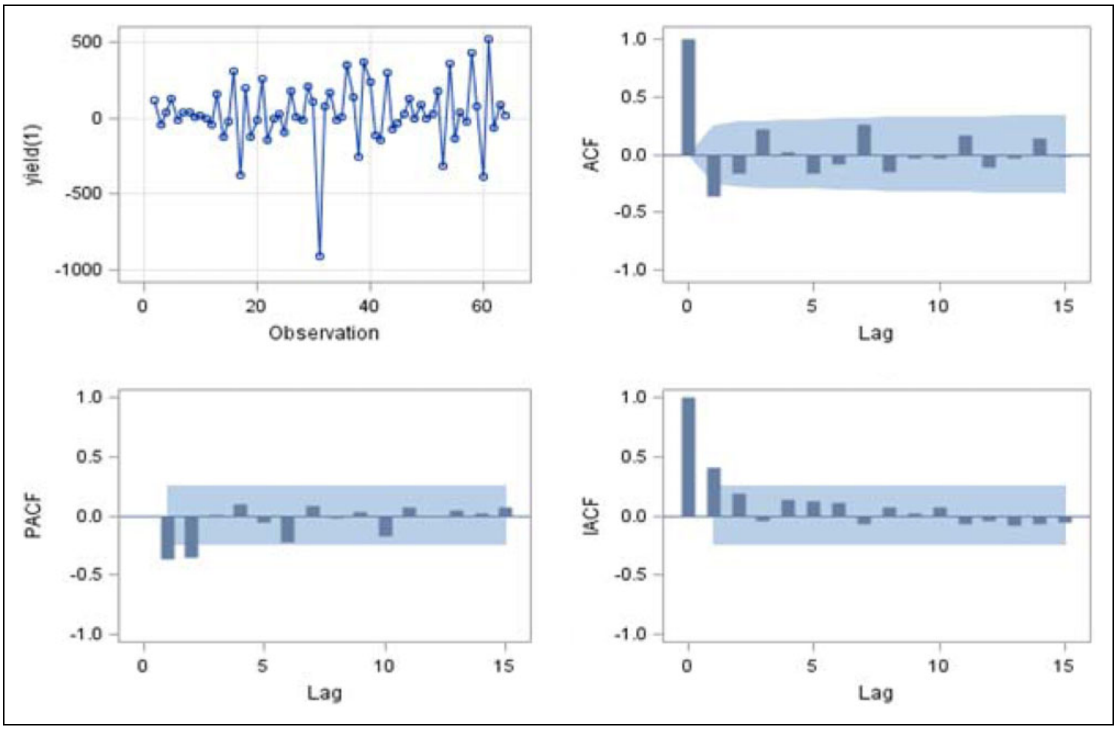

Figure 3. Trend and correlation analyses of differenced (1) series.

our approach, we used the Roulette wheel selection which is based on a randomization process.

Crossover. In crossover, each pair of chromosomes is crossed over to produce two new segments. Usually, offspring inherit some genes from each parent; however, they have their own structures compared with their parents. Crossover operation is not usually applied to all the selected chromosomes. However, the choice is made randomly with a probability of crossover (Pc) being between 0.6 and 1.0. In general, there are three common methods of crossover operation: one-point crossover, multiple-point crossover, and uniform crossover.

Mutation. This is a random search to avoid a premature convergence and is applied to each offspring individually once the crossover operation has been performed. Mutation is a random bit with a small probability Pm (between 0.1 and 0.001) that is randomly selected from the total number of bits from the population matrix.

GA decode. Once the selection, crossover, and mutations are performed, the new offspring are evaluated. In order to do this, the string values should be converted into their equivalent real values. This process is called decoding and is performed by the following equation

$$x = x_{\text{lower}} + \frac{x_{\text{dec}}}{2^n - 1}\left(x_{\text{upper}} - x_{\text{lower}}\right) \qquad (8)$$

where x is gene's real value, x_{dec} is gene's decimal decoded value, x_{lower} is variable lower bound, and x_{upper} is the variable upper bound.

Replacement. Once the new offspring population is evaluated, the parents need to be replaced with the new offspring. The replacement operation is grouped into two main categories, namely, generational and overlapping replacements. In generational replacement, the parent population is replaced by the offspring population except the best individuals in parents, this is also known as nonoverlapping replacement, and in the overlapping replacement, both the offspring and parent population compete to survive into the next generation according to their fitness values.

Termination. Once the convergence criterion is met, such as the maximum number of generations is

Table 3. Testing for stationary in the differenced (1) series.

ADF test statistic				Phillips Perron test statistic			
		Probability				Probability	
Single mean	With trend	Single mean	With trend	Single mean	With trend	Single mean	With trend
−132.40	−326.48	0.0001	0.0001	−75.52	−74.58	0.0006	0.0001

ADF: augmented Dickey–Fuller.

Table 4. Parameter estimation of the ARIMA model by MLE.

MLE					
Parameter	Estimate	SE	*t* Value	*Pr* > |*t*|	Lag
MU	29.04390	12.10634	2.40	0.0164	0
MA1,1	0.50573	0.11133	4.54	<.0001	1

ARIMA: autoregressive integrated moving average; MLE: maximum likelihood estimation; SE: standard error.

Table 5. Autocorrelation of residuals.

autocorrelation check of residuals									
Lag	χ^2	DF	$Pr > \chi^2$	Autocorrelations					
6	6.35	5	0.2739	−0.001	−0.093	0.192	0.020	−0.194	−0.089
12	11.70	11	0.3862	0.174	−0.117	−0.077	−0.006	0.132	−0.059
18	14.72	17	0.6156	0.005	0.144	0.009	−0.076	−0.069	0.061
24	20.24	23	0.6277	0.058	−0.076	0.032	0.191	−0.091	0.024

DF: Dickey–Fuller.

Table 6. Optimal values for the GA parameters.

Population size	250
Selection type	Roulette wheel
Selection rate	60%
Crossover type	Single point
Crossover rate	70%
Mutation rate	0.05
Iteration	27

GA: genetic algorithm.

reached or a desired fitness value is reached, the GA is terminated; if not, the entire algorithm is repeated until the fitness value is reached.

Data description. In this study, all maize yield (kg/ha) data from 1950 to 2013 was collected from agricultural statistics including the Department of Agriculture, Cooperation and Farmers Welfare and Ministry of Agriculture and Farmers Welfare. Data from 1950 to 2003 were used for model building and the remaining decadal data from 2004 to 2013 were used for model validation.

Results and discussion

Summary statistics for Indian maize yield are given in Table 1 and a time series plot is shown in Figure 2. The plot shows that the data set is nonstationary. Figure 2 shows the plot of ACF, PACF, and inverse ACF (IACF) for the actual price series. To validate the stationarity of the series, two tests, namely, ADF test and Philips–Peron test were used (Table 2). The result indicates that the maize yield time series being considered here is nonstationary. Once the series was found to be nonstationary, we made the differencing of order 1 and obtained a resulting stationary series. The differenced series expressed in terms of ACF, PACF, and IACF is given in Figure 3, which shows the series is stationary for the first-order difference. To validate the stationarity of the series, the same two tests were used and results summarized in Table 3. The final model orders based on lowest likelihood ratios, viz., AIC (665.49) and Schwarz's Bayesian criterion (SBC) (669.43), were selected as $p = 0$, $d = 1$, and $q = 1$. Therefore, the resulted model is ARIMA (0, 1, 1) and the parameter estimated by maximum likelihood estimation method is given in Table 4. In ARIMA model building, diagnosis checking is the final step, whereby the autocorrelation of the residuals was evaluated (Table 5) and probability values of the χ^2 test were found to be nonsignificant. We therefore concluded that the chosen model provided a good fit.

Table 7. Parameter estimation of ARIMA by GA.

Statistic	Parameter
MU	27.37
MA(1)	1.54

ARIMA: autoregressive integrated moving average; GA: genetic algorithm.

Table 8. Prediction of maize yield (kg/ha) using the ARIMA and ARIMA-GA approach.

Year	Actual	ARIMA	ARIMA-GA
2004	1907	1933.443	1941.034
2005	1938	1957.489	1985.163
2006	1912	1981.535	2011.942
2007	2335	2005.581	2039.985
2008	2414	2029.626	2067.988
2009	2024	2053.672	2095.975
2010	2542	2077.718	2123.955
2011	2478	2101.764	2151.928
2012	2566	2125.81	2179.892
2013	2583	2149.856	2207.848
MAPE		10.4898	9.91145

ARIMA: autoregressive integrated moving average; GA: genetic algorithm; MAPE: mean absolute percentage error.

Parameter estimation by GA

After the primary classical ARIMA building, an attempt was made to employ the GA for ARIMA parameter estimation. The optimized parameters of GA with regard to minimization of the objective function resulted after several runs. For GA in this problem, the population size was 250 and populations were chosen with a selection rate of 60% based on the Roulette wheel selection method (randomization). The optimum parameter selection for the GA is given in Table 6 and the parameter estimation by the GA is given in Table 7. Once the parameter estimation by GA was complete, the next step was model validation. Data from 1950–2003 were used for model building and the remaining 10-year data (2004–2013) used for model validation, that is, forecasting. The forecast values from the ARIMA and ARIMA-GA approach are given in Table 8. Based on the MAPE (Table 8), one can infer that the ARMA-GA approach can provide reasonable results compared to the other well-known methods as the GA minimizes the error in the parameter estimation and provides excellent results in the model parameter estimation compared to classical methods. This is also important for future maize forecasting studies. This approach could be further extended using other machine learning techniques for varying the AR and MA orders so that practical validity of the model could be well established.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Agricultural situation in India (2015) Directorate of Economics and Statistics, Department of Agriculture and Cooperation, Ministry of Agriculture and Farmers Welfare, GOI.

Agricultural statistics at a glance (2014) DES, Department of Agriculture, cooperation and Farmers welfare, Ministry of Agriculture and Farmers welfare, GOI.

Box GEP and Jenkins GM (1970) *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Box GEP, Jenkins GM, and Reinsel GC (1994) *Time Series Analysis: Forecasting and Control*, 3rd ed. San Francisco: Holden-Day.

Choi HS, Schneider UA, Rasche L, et al. (2015) Potential effects of perfect seasonal climate forecasting on agricultural markets, welfare and land use: a case study of Spain. *Agricultural Systems* 133: 177–189.

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

Haupt SE and Haupt RL (2004) *Practical Genetic Algorithms*. UK: Wiley.

Holland J (1975) *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: The University of Michigan Press.

Jambhulkar NN (2013) Forecasting of rice production in Punjab using ARIMA model. *International Journal of Scientific Research* 2(8): 1–2.

Karim R, Awala A, and Akhter M (2010) Forecasting of wheat production in Bangladesh. *Bangladesh Journal of Agricultural Research* 35(1): 17–28.

Naveena K, Rathod S, Shukla G, et al. (2014) Forecasting of coconut production in India: a suitable time series model. *International Journal of Agricultural Engineering* 7(1): 190–193.

Parviz L, Kholghi M, and Hoorfar A (2010) A comparison of the efficiency of parameter estimation methods in the context of streamflow forecasting. *Journal of Agricultural Science and Technology* 12: 47–60.

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

Report of India maize summit, New Delhi, 9–10 April 2015.

Rolf S, Pravez J, and Urfer W (1997) Model identification and parameter estimation of ARMA models by means of evolutionary algorithms. *Computational Intelligence for Financial Engineering* 23: 237–243.

Sarika MAI and Chattopadhyay C (2011) Modelling and forecasting of pigeonpea (*Cajanus cajan*) production using autoregressive integrated moving average methodology. *Indian Journal of Agricultural Sciences* 81(6): 520–523.

Tahir A and Habib N (2013) Forecasting of maize area and production in Pakistan. *ESCI Journal of Crop Production* 2(2): 44–48.

Zaer SA, Alsmadi MK, Alsmadi AM, et al. (2012) ARMA model rder and parameter estimation using genetic algorithms. *Mathematical and Computer Modelling of Dynamical Sys-tems: Methods, Tools and Applications in Engineering and Related Sciences* 18(2): 201–221.