Autoregressive Integrated Moving Average (ARIMA) ModelWith Genetic Algorithm to Forecast the Chilli and Turmeric Productions in India

ABSTRACT

Aims: India holds the distinction of being the foremost producer of spices globally and has been long-run history in spice export. The quantity of Indian spice exports increased by 37% with $ 4.1 billionworth in 2021. With that, dried chilli, cumin, and turmeric alone contributed 44% of export value ($ 1.8 billion). Forecasting the production of major spices are key for exportsand plays an essential role in supporting and achieving the target of $10 billion in exports by 2027.
Data Source: The time series data of chilli and turmeric production data in India from 1970-2020 periods was collected from Indiastat.
Methodology: The present study sought to forecast the production of chilli and turmeric in India using the ARIMA model and their parameters are estimated by stochastic optimization techniques (genetic algorithm). The parametersare estimated by minimizingthe Mean Absolute Percentage Error(MAPE). Finally,ARIMA and ARIMA_GA modelswerecompared based on their predictive ability.
Results: The Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were254.01,11.32 (chilli) and 185.73, 15.24 (turmeric) for testing set of ARIMA_GA model which is lower than the fitted ARIMA model..
Conclusion: This work has shown that ARIMA_GA (2,1,1)has been thebest model to forecast the chilli and turmeric production in India. ARIMA_GA model will cope with parsimony and convergence of likelihood function to global optimum problems.ThereforeARIMA with GA will able to model the complexity and uncertainty of the data.

Keywords: Maximum likelihood estimate, ARIMA, Genetic algorithm and MAPE

1. INTRODUCTION

Indian spice and spice products have demand in over 180 countries due to their unique aroma, taste and medicinal benefits [1]. chilli holds the second largest area and production among all the spices in India. Turmeric ranks fourth in production and sixth in the area under cultivation. Moreover, chilli and turmeric were the most exported spices to China, the USA, Bangladesh, Thailand, UAE, Sri Lanka, Malaysia, the UK, Indonesia, and Germany during 2020-21. The contribution of chilli and turmeric in export is 36% and 10%, respectively [2.

ARIMA model is commonly used to forecast the time-dependent univariate time series model for agricultural production [3]. Generally, Box-Jenkins Autoregressive Integrated ovingAverage (ARIMA) time-series strategyhas been considered for ARIMA model fitting [4]. The parameters of ARMA (p, q) models were estimated by the maximum likelihood function for given observed time series values [5]. Stationary of the time series is the principal condition for model fitting, which is defined as joint distributions of time series and time-shifted vectors are the same [6].Biswas and Bhattacharyya [7]used Box-Jenkins ARIMA model to forecast the important pulse crops of Odisha and their orders haveidentifiedbased onAutocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) at various lags. Forecasted the area and production of rice in West Bengal using ARIMA model, among the series of ARIMA models, the best fitted model was selected based on the AIC, SBC and MAPE criteria [8].

Inboth maximum likelihood and least-square methods,the corresponding function must be established and optimized. The optimization should be done by numerical methods which would not converge to global optimum in case of complexity. Also, the order of the ARIMA model used to select as low as possible due to parsimony effects[9]. The genetic algorithm originated from Darwin's evolutionary principles that may be used to search for the optimal solution to a problem [10]. Parviz et al., [11] indicated genetic algorithm was more appropriate than other methods such as conditional likelihood

and unconditional likelihood methods of estimation. Becauseofa high convergence speed to the global optimum for the complexity of the model. GA has been used to identify and estimate the parameters ofARIMA model [12].

The order of the ARIMA model and their estimation obtained by GA would increasedthe accuracy of prediction [13].Rathod et al.,[14] utilized ARIMA-GA to forecast the maize production in India.Alquraish etal.,[15] showedthat the ARIMA-GA model performed better than the hidden Markov model-GA. novel ARIMA-GA-ANN to forecast the standard precipitation index (SPI) in the Bisha Valley, Saudi Arabia.Solar radiation prediction by GA based model has better than extreme gradient boosting [16]. Therefore, employing the GA for parameter estimation in the ARIMA model to forecast the major spices will enhance the forecast accuracy, for crucial role in policy making and decidingthe nation's income through export returns.

2. MATERIALS AND METHODS

2.1 Stationarity and Autocorrelation of time series models

The time series data of chilli and turmeric production data in India from 1970-2020 periods weredivided into training (1970-2012) and testing sets (2013-2020) at the ratio of 85 % and 15 %. The data hadn't shown any seasonality so thenonseasonal ARIMA (p,d,q) model is used and their generalized is given in equation 1 [18].

$$\emptyset_p(B)(1-B)^d Y_t = \theta_q(B)\epsilon_t$$

$\emptyset_p(B) = (1 - \emptyset_1 Y_{t-1} - \emptyset_2 Y_{t-2} - \cdots - \emptyset_p Y_{t-p})$ (p-order AR operator)

$\theta_q(B) = (1 - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q})$(q-order MA operator) (1)

Where, $(1-B)^d$ – d order differencing operator.Generally, time series indicates the existence of some form of dependence between the observations. The dependency and their structure are determined by the autocorrelation function and stationarity process Augmented Dickey Fuller(ADF) test is used to test a null hypothesis that the time series is unit root (not stationary) against the alternative hypothesis is that the time series is stationary. The order of AR(p), and MA(q) parameters are determined by the visualization of Autocorrelation Function (ACF)and Partial Autocorrelation Function (PACF)while best fitted model is selected based on Akaike Information Criterion (AIC)and Bayesian Information Criterion (BIC)

2.2 Estimation of parameter

Often the order of AR(p) and MA(q) parameters are estimated by the maximum likelihood method or least square method. In the case of the least square method minimize the conditional sum of square $S_c(\emptyset, \theta) = \sum_{t=m}^{n} e_t^2$ with respect to parameters. The likelihood function (L) is a joint distribution function of unknown parameters of $\emptyset$, θ and $\sigma^2$for given observations $Y_1,\ldots\ldots,Y_n$. The maximum likelihood function of ARMA (p,q) is given in equation 3 [6].

$$L(\emptyset, \theta, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^n r_0 \ldots\ldots r_n}} exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\frac{(Y_j - \hat{Y})^2}{r_{j-1}}\right\} \quad (2)$$

$$\hat{Y}_{n+1} = \emptyset_1 Y_n + \cdots + \emptyset_p Y_{n+1-p} + \sum_{j=1}^{q}\theta_{nj}\left(Y_{n+1-j} - \hat{Y}_{n+1-j}\right) \quad n \geq m \ (m = \max(p,q)) \quad (3)$$

$E(Y_{n+1} - \hat{Y}_{n+1})^2 = \sigma^2 r_n$, is determined recursively.

For any fixed *p* and *q*it is clear that the estimates of the $\emptyset$, θ parameters are that minimize $-2lnL(\emptyset, \theta, \sigma^2)$i.e., the maximum likelihood estimators.

2.3 Genetic algorithm

The genetic algorithm is a stochastic search and optimization procedure motivated by the principles of genetics and natural selection [10]. It combines Charles Darwin's principles of "Natural selection" and "Survival of the fittest" with a computer-constructed evolution mechanism to select better species from the original population. This is done by random exchange of information among them, expecting superior offspring.

In the GAs, a population of possible solutions is evaluated to estimate the best solution. GAs is based on three main concepts viz., reproduction, evaluation and selection. Genetic reproduction is performed using two basic genetic operators viz., crossover and mutation. The evaluation is performed using 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.

The construction of the GAs for any problem can be separated into five distinct tasks: (1) representing genetically potential problem solutions; (2) creating an initial population of solutions; (3) designing genetic operators; (4) implementing the fitness functions; and (5) setting the system parameters, including population size, probabilities with which genetic operators are applied and so on. Each of the mentioned components greatly affects the solution obtained as well as the performance of the GA[14].

2.4 Genetic Algorithm – ARIMA model

Considering the advantages of genetic algorithm which has been applied in the ARIMA model parameter estimation. The steps involved in ARIMA_GA models is described in the following subsections.

2.4.1Initialization

Identifying the order of the ARIMA model and the initializationof its parameters simultaneously are the main characteristicsofARIMA model fitting. The stationarity of the model is assessed by unit root test

or ADF test and the order of the model is determined with the lowest value of both AIC and BIC criteria as $n\ln(L) + 2(p + q +)$ and $n\ \ln(L) + 2(p + q +)\ln(n)$.
The search space limits of AR(p) and MA(q) coefficient for stationarity time series areconsidered as -0.999 to 0.999 owing to causality and invertibility process and for constant is 2 or 3 times of its original ARIMA(p,d,q) models' standard errors. Thiswould create aset of possible solutions (called population) with specific size is referred to as population size and each solution is a chromosome which would be binary or real valued.

2.4.2 Fitnessevaluation

The objective function should be defined in terms of fitness function for evaluation. The best solution (AR(p) and MA(q) coefficient) is obtained by minimizing the given fitness function at the specified rate of generations from the population [14].

$$Fitness = \frac{1}{1+MAPE}$$
(4)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_t - \hat{Y}_t}{Y_t} \quad (5)$$

Mean Absolute Percentage Error (MAPE) as the fitness function. Where $Y_t$is the actual value and $\hat{Y}_t$ is the estimated value.

2.4.3Selection

For creating new offspring (solution) for the next generation, set of best solutions are chosen from the current population or sets of possible solution space based on the fitness function. The commonly used method is Roulette wheel selection.Elitism is the Copy the best solution to creating a new population before applying crossover and mutation. Forces GAs to retain some number of the best individuals at each generation.

2.4.4Crossover

In a crossover, each pair of chromosomes is crossed over to produce two new segments. Usually, offspring inherit some genes from each parent. The crossover is made randomly with a probability of crossover (P_c) being between 0.6 and 1.0.

2.4.5Mutation

This is a random search to avoid premature convergence and is applied to each offspring individually once the crossover operation has been performed. The mutation is a random bit with a small probability P_m (between 0.1 and 0.001) that is randomly selected from the total number of bits from the population matrix.The selected parameter values for the genetic algorithm are given in Table 1.

Table 1. Parameters selected for genetic algorithm

GA parameters	Values
Population size	100 -150
Number of generations	500
Elitism	5
Crossover probability	0.8
Mutation probability	0.2

2.4.6The Portmanteau tests

For large *n,* the sample autocorrelations of an iid sequence Y_1, ……, $Y_n$with finite variance are approximately iid with distribution N(0, 1). Hence, if Y_1, ……, $Y_n$is a realization of such an iid sequence, about 95% of the sample autocorrelations should fall between the bounds $\pm 1.96/\sqrt{n}$which is the test by Ljung-Box statistic, $Q = n(n+2)\sum_{j=1}^{h}\hat{\rho}_j^2\ /n-j$ (6)

2.5**Model evaluation criteria**

Criteria that are used to make the comparison of forecasting ability among different models are Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) [18]. These errors are on the same scale as the data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} \quad (7)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - \hat{Y}_t|}{Y_t} \times 100 \quad (8)$$

Where, Y_t is the actual value of chilli and turmeric time series, $\hat{Y}_t$ is the predicted value and n is the number of observations.

3. RESULTS AND DISCUSSION

For training sets,the stationarity of the data is tested by the ADF test which stated actual times series of chilli and turmeric is non-stationary due to their P value is greater than 0.01.Therefore, both chilli and turmeric series required one difference.After the 1st difference, P value is less than and equal to 0.01 would confirm that the series become stationary (Table 1). which can be visually noticed in ACF and PACF plots. The sin wave of ACF plots of actual time series indicated that both chilli and turmeric series were non-stationary (Figure1&2).

Table2. Testing the Stationarity of time series by ADF test

	Lags	Statistic	P value
Chilli	3	-3.46	0.06
Turmeric		-1.87	0.62
	After first difference		
Chilli	3	-5.00	0.01**
Turmeric		-6.08	0.01**

** 1% significant

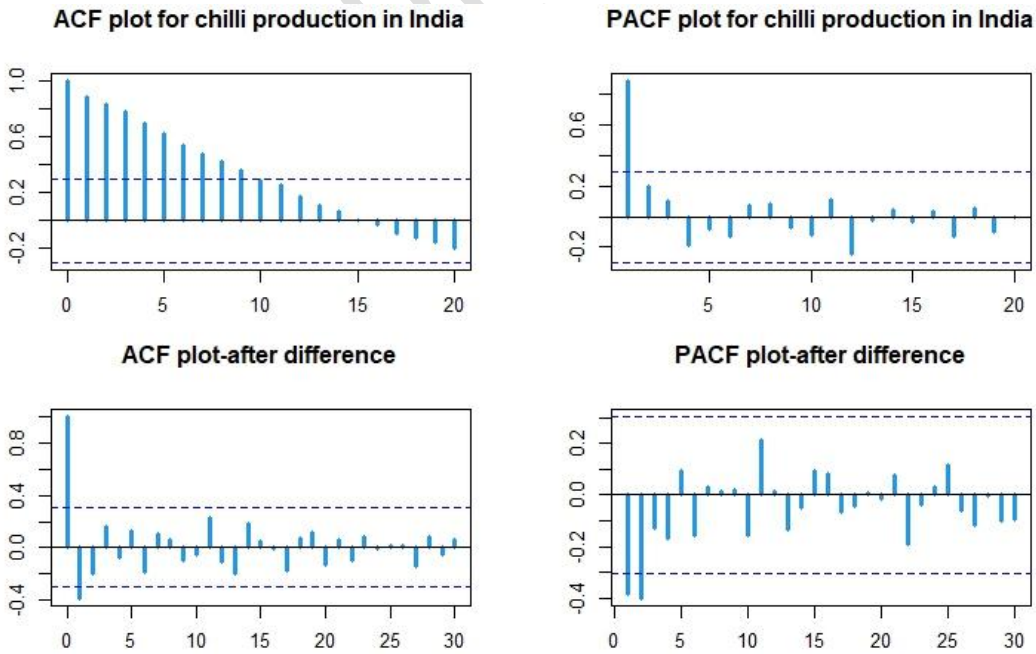

Fig. 1ACF and PACF plot of chilli production time series in India

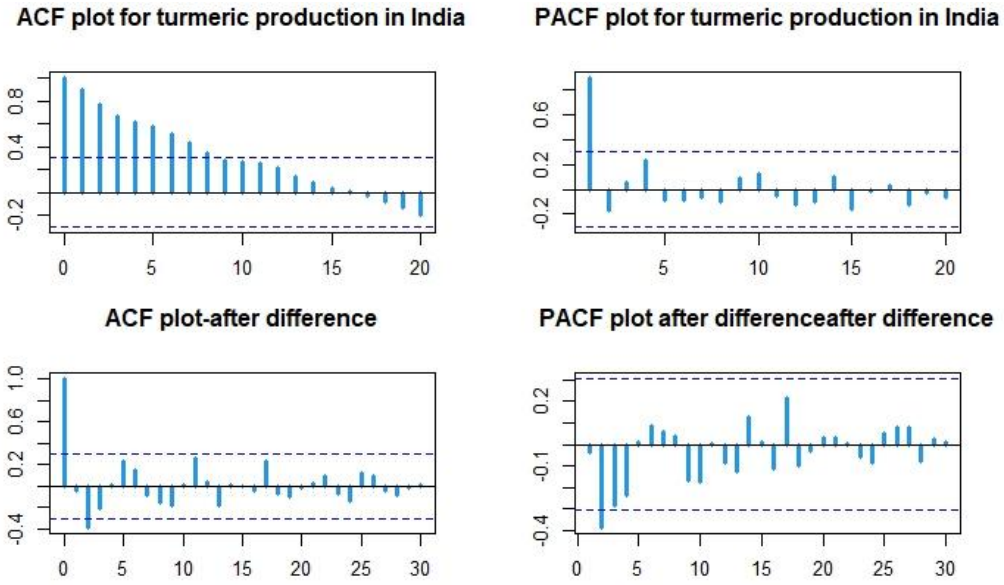

Fig. 2ACF and PACF plot of turmeric production time series in India

The model, ARIMA (2,1,1) hasbeen selected as best fitted based on the lowest value of AIC and BIC criteria for both crops. The coefficients of the ARIMA model were estimated by maximizing the likelihood function (L) and in the case of ARIMA_GA the coefficients have been determined by stochastic search from the population with 100-150 size subject to minimizing the mean absolute percentage error at 9.99% for chilli and 13.34% for turmeric over 500 generations. Almost the best solution converged after the 10^{th} and 50^{th} generations for chilli and turmeric (Figures 3and4). The estimated value of ARIMA parameters with their standard errors and ARIMA_GA parameters are given in Table 3.

Table 3. Estimated coefficient values of fitted ARIMA and ARIMA_GA models

	Models	AIC	BIC	AR(1)	AR(2)	MA(1)	Constant
Chilli	ARIMA (2,1,1)	515.73	524.42	-0.014 (0.24)	-0.19 (0.19)	-0.67 * (0.23)	22.41* (4.52)
	ARIMA_GA (2,1,1)			-0.114	-0.235	-0.351	23.553
Turmeric	ARIMA (2,1,1)	503.05	511.74	0.36 (0.21)	-0.44* (0.15)	-0.57* (0.20)	22.86* (5.41)
	ARIMA_GA (2,1,1)			-0.071	-0.461	-0.062	18.165

* 5% significant, parenthesis : Standard Error,AR(1) : Autoregressive order and MA(1) : Moving average order

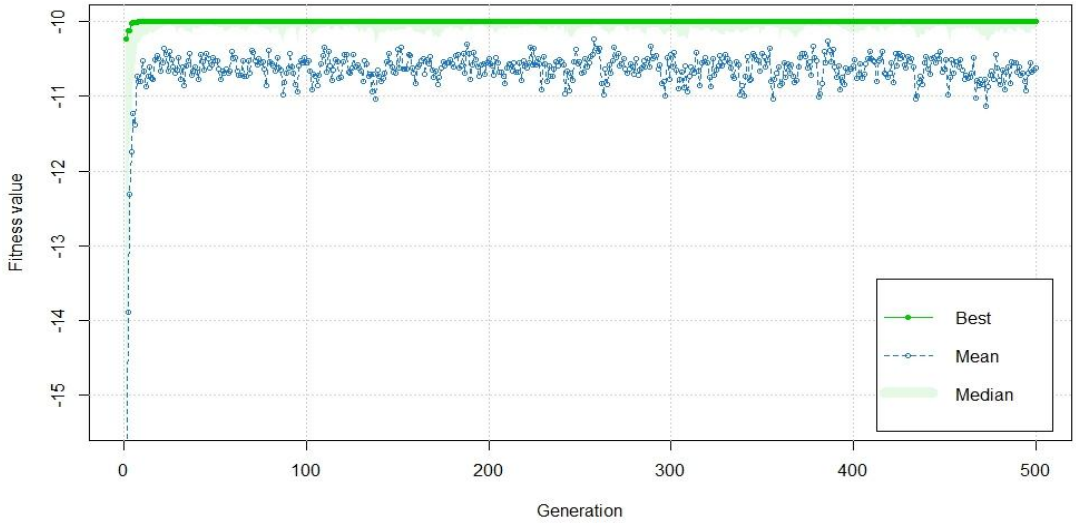

Fig. 3Fitness function values (MAPE) over generations of chilli production

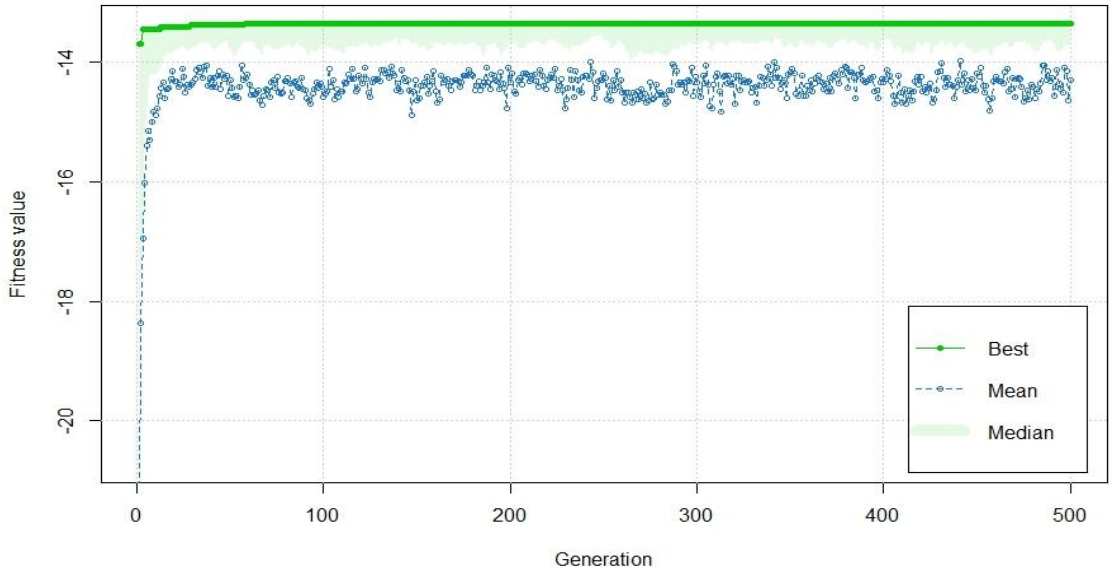

Fig. 4Fitness function values (MAPE) over generations of turmeric production

The predictive ability of fitted ARIMA and ARIMA_GA models was assessed by root mean square error which is a unit measure (measured unit of data) and mean absolute percentage error. The ARIMA_GA model performed better in both the training and testing data sets for chilli and turmeric productionowing tocomparatively less valuein RMSE and MAPE measures(Table4). The P value of theLjung-Box test was greater than 1 %, which might accept the null hypothesis that the white noise of ARIMA_GA models' residuals (Table5). The actual and fitted values of the best fitted model of chilli and turmeric in India areshown in Figures 5 and 6.

Table4.Comparison ofthe predictive ability of fitted models

	ARIMA				**ARIMA_GA**			
	Training		Testing		Training		Testing	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Chilli	106.83	10.61	263.60	11.41	104.49	9.99	254.01	11.32
Turmeric	95.22	13.91	189.80	16.31	94.06	13.34	185.73	15.24

Table 5. Ljung-Box test to check the white noise of residuals

	Lags	**ARIMA_GA**	
		Statistic	P value
Chilli	13	6.96	0.54
Turmeric		11.70	0.16

The forecasting using ARIMA (p,d,q) is based on the lagged or past values of chilli and turmeric production along with stochastic error terms which explain the probabilistic or stochastic nature of the observation over a period of a particular time(economic time series). The probability value in the ADF test of chilli and turmeric productions was higher than 0.05 which supportsto acceptance null hypothesis which means the non-stationary of the series. But the P value of the differenced series is less than 0.01 indicating the stationarity [18]. The fitted ARIMA model for chilli production, ARIMA (2,1,1), differs from the findings of Padmanaban et al.[17], whose model ARIMA (0,1,1) was identified as the best for chilli production in India from 1970 to 2012. This variance can be attributed to the different durations of data analysed in the respective studies.Though the ARIMA model has captured the past linear relationship effectively in the system and their parameters are estimated by maximum likelihood estimation, the estimated parameters are unstable and non-significant when the data has outliers or leverage points andparsimony. Therefore, Rathod et al.,[14] applied a stochastic global search algorithm with respect to Darwin's natural selection (genetic algorithm) for parameter estimation of ARIMA and interpreted ARIMA_GA performed better than normal ARIMA model for maize production. Similar resultscan also be obtained for chilli and turmeric production.The residuals of fitted ARIMA_GA models for both chilli and turmeric production haveefficiently supported the Ljung-

Box test's null hypothesis which would ensure the reliability of the fitted model for forecasting future values, the result is consistent with Abbasi et al.,[13].

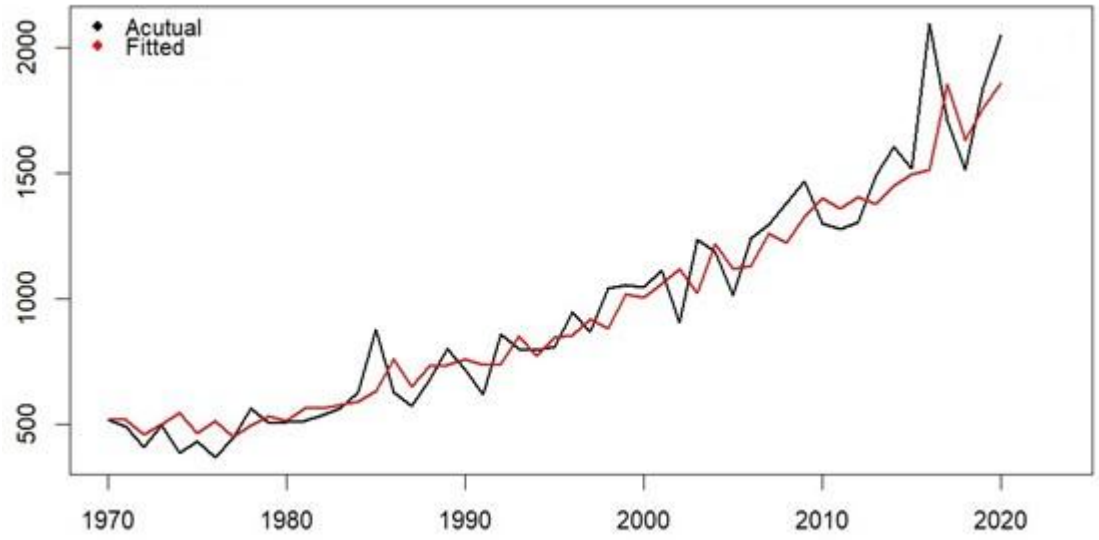

Fig. 5Actual vs fitted value and forecasted of chilli production in India

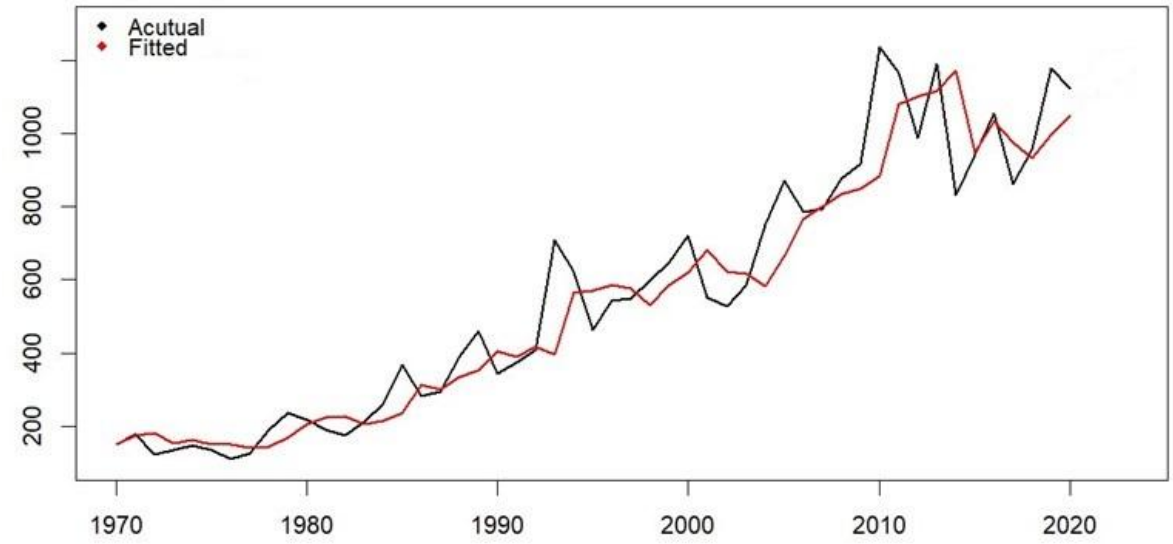

Fig. 6: Actual vs fitted value and forecasted turmeric production in India

4. CONCLUSION

Chilli and turmeric are among the major spices exported from India, with turmeric being particularly notable for its medicinal properties. During the first half of the COVID-19 crisis, turmeric accounted for 42% of the total volume of spice exports from India. The present study focused on forecasting the time series of chilli and turmeric. In general, forecasting time series is not deterministic due to random components. However, if the random component is stationary, it is possible to develop sound techniques to forecast its future values. Here, the stationarity of the time series was checked by ADF test and the parameters of the ARIMA (2,1,1) model for both chilli and turmeric time series, estimated by maximum likelihood method and genetic algorithm were compared in both training and testing data sets based on the lowest value of RMSE and MAPE. The results indicated that the genetic algorithm demonstrated an improvement in prediction ability in both the training and testing datasets compared to the maximum likelihood method.

REFERENCE

1. India Brand Equity of Foundation (IBEF). Indian Spices, Spices Manufacturers and Exporters in India – IBEF, 2023. http://www.ibef.org/exports/spice-industry-indias.
2. Directorate General of Commercial Intelligence and Statistics (DGCI&S). (2020). Available: https://www.indianspices.com/sites/default/files/Major_item_wise_Export_2020.pdf.
3. Mohammad N, Islam M. A, Rahman M, Mahboob MG. Forecasting of maize production in bangladesh using time series data: The Bangladesh Journal of Agricultural Economics, 2022; 43(2): 18-32.
4. Box GEP, Jenkins GM. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day; 1970.
5. Hamjah MA. Forecasting major fruit crops productions in Bangladesh using Box-Jenkins ARIMA model: Journal of Economics and Sustainable Development. 2014; 5(7): 96-107.
6. Brockwell PJ, Davis RA. Introduction to time series and forecasting. Springer-Verlag: New York; 1996. pp.43-75
7. Dash A, Mahapatra SK. Using ARIMA model for yield forecasting of important pulse crops of Odisha, India: Amazonian Journal of Plant Research. 2020; 4(3):646-659. Available:10.26545/ajpr.2020.b00073x.
8. Biswas R. Bhattacharyya B. ARIMA modeling to forecast area and production of rice in West Bengal: Journal of Crop and Weed. 2013; 9(2): 26-31.
9. Rolf S, Pravez, J. Urfer W. Model identification and parameter estimation of ARMA models by means of evolutionary algorithms:Computational Intelligence for Financial Engineering. 1997; 23: 237-243.
10. Holland J. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: The University of Michigan Press; 1975.
11. Parviz L, Kholghi M, Hoorfar A. A comparison of the efficiency of parameter estimation methods in the context of streamflow forecasting: Journal of Agricultural Science and Technology. 2010; 12: 47-60.
12. Zaer SA, Alsmadi MK, Alsmadi AM. ARMA model order and parameter estimation using genetic algorithms: Mathematical and Computer Modelling of Dynamical Systems: Methods, Tools and Applications in Engineering and Related Sciences. 2012; 18(2): 201-221.
13. Abbasi A, Khalili K, Behmanesh J, Shirzad A. Estimation of ARIMA model parameters for drought prediction using the genetic algorithm: Arabian Journal of Geosciences. 2021; 14(10): 841.
14. Rathod S, Singh KN, Arya P, Ray M, Mukherjee A, Sinha K, Kumar P, Shekhawat RS. Forecasting maize yield using ARIMA-genetic algorithm approach: Outlook on Agriculture. 2017; 46(4): 265-271.
15. Alquraish M, Abuhasel K, Alqahtani S, Khadr M. SPI-based hybrid hidden Markov-GA, ARIMA-GA, and ARIMA-GA-ANN models for meteorological drought forecasting: Sustainability. 2021; 13.
16. Gunasekaran V, Kovi KK, Arja S, Chimata R. Solar irradiation forecasting using genetic algorithms: ArXiv preprint *arXiv:2106.13956*. 2021; Available: 10.48550/arXiv.2106.13956.
17. Padmanaban K, Sahu PK, Narsimhaiah L. Production performance of chilli in India- a statistical approach: Advances in Life Sciences. 2016; 5(10): 4191-4200.
18. Dheer P. (2019). Time series modelling for forecasting of food grain production and productivity of India: Journal of Pharmacognosy and Phytochemistry. 2019; 8(3): 476-482.