Original Research Article

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 export increased by 37% during 2021. India exported $ 4.1 billion worth of spices and their products during 2021-22. From this share, dried chilli, cumin, and turmeric alone contributed by 44% ($ 1.8 billion). Forecasting the production of major spices, which are key for exports, 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: Sometimes autoregressive integrated moving average (ARIMA) models facing parsimony and convergence of likelihood function to global optimum. Thus, 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). Instead of minimizing the residual sum of squares, the mean absolute percentage error is minimized. Both ARIMA and ARIMA_GA model was compared based on their predictive ability.
Results: Root mean square error (RMSE) and mean absolute percentage error (MAPE) was observed to be minimum for ARIMA_GA in both training and testing data sets.
Conclusion:ARIMA_GA is found to best for forecasting the chilli and turmeric production in India while the parameters of ARIMA are stochastically estimated using a genetic algorithm. It would give the global converges to the function which has the complexity in moving average term.

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

1. INTRODUCTION

Indian spice and spice products have demand in over 160 countries due to their unique aroma, tasteand medicinal benefits. 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, USA, Bangladesh, Thailand, UAE, Sri Lanka, Malaysia, UK, Indonesia, and Germany during 2020-21. The contribution of chilli and turmeric in export is 36% and 10%, respectively [8].

~~The~~ ARIMA model is commonly used to forecast the time-dependent univariate time series model for agricultural production. Generally, Box-Jenkins Autoregressive Iintegrated Mmoving Aaverage (ARIMA) time-series strategyhas been considered for ARIMA model fitting [4]. 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 [5].Box-Jenkins autoregressive integrated moving average time-series methodology to forecast the important pulse crops of Odisha and orders of the different ARIMA models hasidentifiedbased onautocorrelation function (ACF) and Partial autocorrelation unction (PACF) at various lags [6]. Forecasted the area and production of rice in West Bengal using ARIMA model, among the series of ARIMA model, best model was selected based on the AIC, SBC and MAPE criteria [3].

The parameters of ARMA (p, q) models were estimated by maximizing the maximum likelihood function for given observed time series values [10]. Inboth maximum likelihood and least-square methods, corresponding function must be established and optimized. The optimization should be done by numerical methods which would not converges to global optimum in case of complexity. Due to principles of parsimony, keeping the parameters as low as possible in ARIMA models [15]. The

Comment [AA1]: Model

Comment [AA2]: During might be changed to 'in'.

Comment [AA3]: Both sentences can be merged together

Comment [AA4]: Another conjunction might be better here

Comment [AA5]: 44 % of what?

Comment [AA6]: It is an incomplete sentence. Recast

Comment [AA7]: State only what you did, no need for instead …

Comment [AA8]: You need to state the specific values of your results

Comment [AA9]: The concise result of the ARIMA and GA methods should be included.

Comment [AA10]: Recast the sentence. You can start with
This work has shown ……

Comment [AA11]: reference

Comment [AA12]: Let your references be in order [1], [2], …

Comment [AA13]: ref

Comment [AA14]: Use correct reference number

Comment [AA15]: Recast the sentence

Comment [AA16]: The sentence should be taken to immediately after sentence 1 para 2

Comment [AA17]: Incomplete sentence

genetic algorithm came from Darwin's evolutionary principles that may be used to search for the optimal solution to a problem [11]. Parviz et al., [13] indicated genetic algorithm was more appropriate than other methods such as conditional likelihood and unconditional likelihood and genetic algorithm (GA) has a high convergence speed to global optimum for complexity of the model. GA has been used to identify and estimate the parameters ofARIMA~~Arima~~ model [16].

The order of ARIMA model and their estimation obtained by GA which would increasing the accuracy of prediction [1].ARIMA-GA approach to forecast the maize production in India [14].Alquraish etal.,[2] shown that ARIMA-GA model performed better than the hidden Markov model-GA., ~~and a~~ novel ARIMA-GA-ANN to forecast the standard precipitation index (SPI) in the Bisha Valley, Saudi Arabia.Improved accuracy by GA for solar radiation prediction than extreme gradient boosting [9].Therefore, employ the GA for parameter estimation in ARIMA model to forecasting the major spices has enhance the forecast accuracy, for crucial role in policy making and decide the nations income through export returns.

2. MATERIALS AND METHODS

2.1 Stationarity and Autocorrelation of time series models

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. Generally, a nonseasonal ARIMA model is denoted as ARIMA (p, d, q) and expressed as

Comment [AA18]: Another word will be better here. Such as originated, ...

Comment [AA19]: Break the sentence

Comment [AA20]: Incomplete sentence. You can start the sentence with the Author;s name.

Comment [AA21]: This sentence does not link with the first part

Comment [AA22]: Recast the sentence

Comment [AA23]: Check for grammatical error and the use of the past tense.

The sentence might start with
Researches has shown that the use of GA for parameter estimation in ARIMA model ...

Comment [AA24]: There are different types of ARIMA models. Justify the usage of nonseasonal ARIMA model and include reference.

Comment [AA25]: Reference

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

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

$(1-B)^d$ – d order differencing operator.

Include the definition of terms

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 ACF and PACFwhile~~and~~ best fitted model is selected based on AIC and BIC criteria.

2.2 Estimation of parameter

Once the identified 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,……,Y_n. The maximum likelihood function of ARMA (p,q) is given by [reference]

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

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

$E(Y_{n+1} - \hat{Y}_{n+1})^2 = \sigma^2 r_n$, 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 [11]. 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 [ref].

2.4.1Initialization

Identifying the order of the ARIMA model and the initialization its parameters simultaneously are the main characteristic in ARIMA– model fitting. The stationarity of the model is assessed by unit root test or ADF test and order of the model is determined with 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 is consider 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. Which would create the sets of possible solution (called

Comment [AA26]: Include equation number

Formatted: Tab stops: 4.29", Left

Comment [AA27]: Start a sentence with the full meaning. Abbreviation can be in bracket

Comment [AA28]: Include the full definition of the terms, ACF and others

Comment [AA29]: Incomplete sentence

Comment [AA30]: Include definition of terms

Comment [AA31]: A section heading should come before this as 2.4 GA- ARIMA in full. Introduce it before going to this is a subsection under 2.4, it should be 2.4.1

population) with specific size is referred as population size and each solution is called as chromosome which would be binary or real valued.

2.4.2 ~~2.5~~Fitness evaluation

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 [ref].

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

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

Include equation number

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

2.4.3~~6~~Selection

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.

2.4.4~~7~~Crossover

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.

2.4.5~~8~~Mutation

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 parameter used in GA is given below:

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

Comment [AA32]: Include table title

2.4.6~~9~~The Portmanteau tests

For large *n,* the sample autocorrelations of an iid sequence $Y_1, \ldots\ldots, Y_n$ with finite variance are approximately iid with distribution N(0, 1). Hence, if $Y_1, \ldots\ldots, Y_n$ 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,

$$Q = n(n+2)\sum_{j=1}^{h} \hat{\rho}_j^2 / n - j$$

You need to include the method of RSME and other statistical analysis performed. Also the performance evaluation method.

Formatted: Left

3. RESULTS AND DISCUSSION

The time series data of chilli and turmeric production data in India from 1970-2020 periods were divided into training and testing sets at the ratio of 85 % and 15 %. For training sets stationarity of the data is tested by the ADF test and 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 (Fig. 1&2).

Comment [AA33]: It should be in methodology

Table~~2~~1. 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

Comment [AA34]: The table is not well discussed. From my observation of the P value in the first segment is insignificant. You need to discuss the reason and compare your work with available literatures.

The order of the model ARIMA (2,1,1) has been selected at 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 10th and 50th generations for chilli and turmeric (Fig 3&4). The estimated value of ARIMA parameters with their standard errors and ARIMA_GA parameters are given in Table ~~3~~2.

Comment [AA35]: How did you obtain this. If possible, include the method in your methodology

Comment [AA36]: This should be Figure 1 and 2

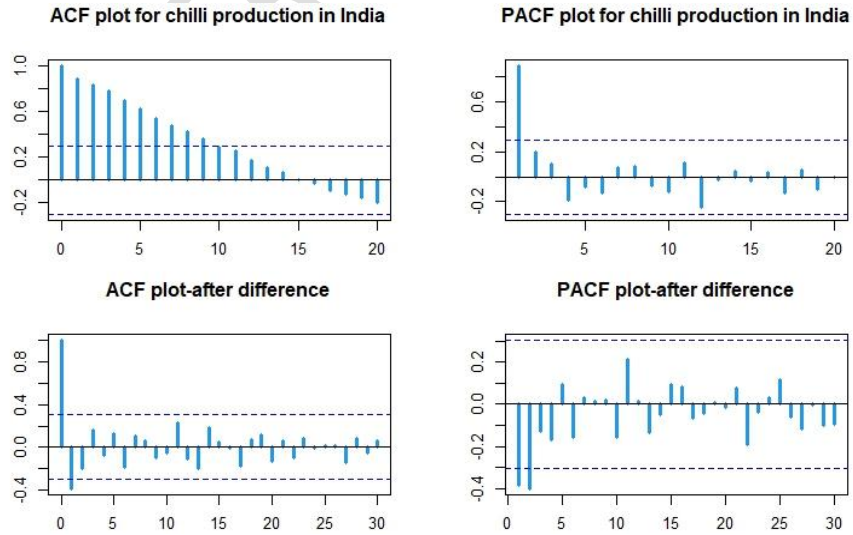

Fig. 1ACF and PACF plot of chilli production time series in India

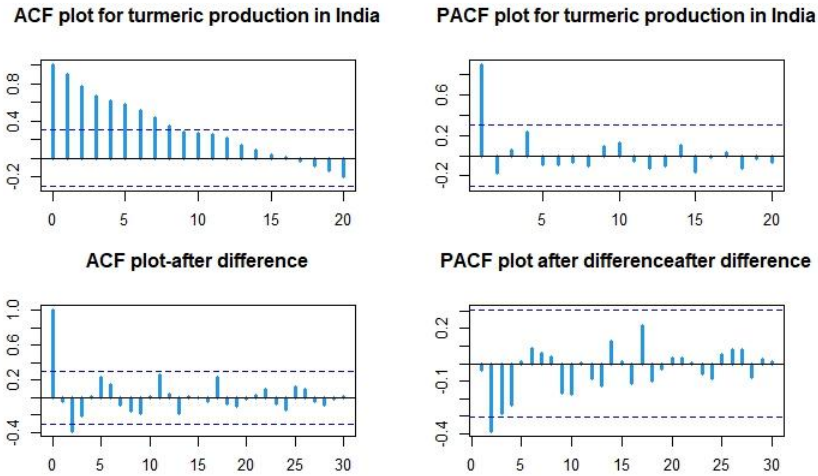

Fig. 2ACF and PACF plot of turmeric production time series in India

Table 2. Estimated coefficient values of fitted ARIMA and ARIMA_GA models

	Models	AIC	BIC	AR1	AR2	Ma1	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

Include definition of terms

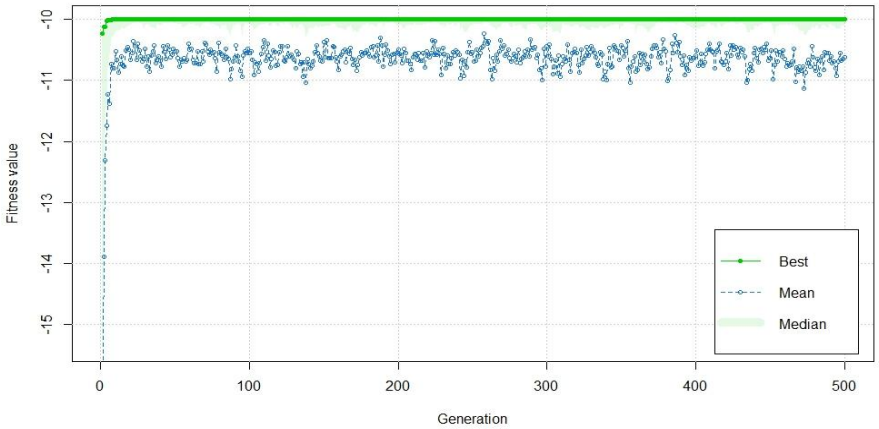

Fig. 3Fitness function values (MAPE) over generations of chilli production

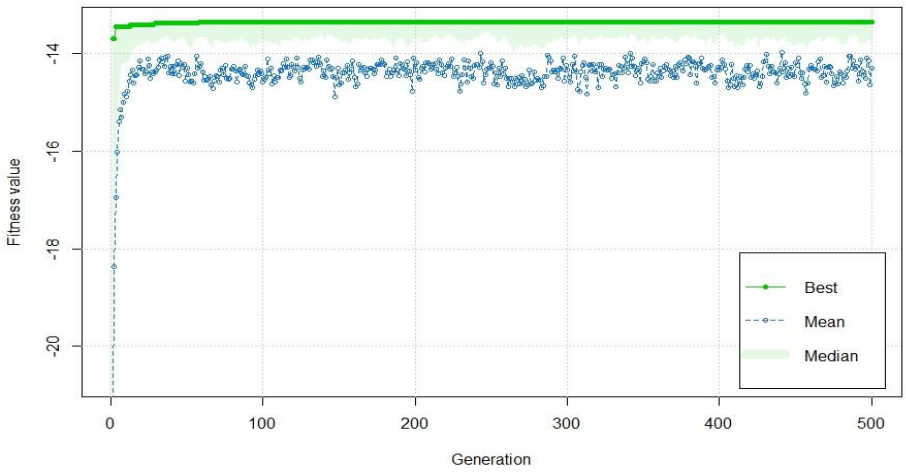

Fig. 4Fitness function values (MAPE) over generations of turmeric production

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 thetraining (1970-2012) and testing data sets (2013-2020) for chilli and turmeric productionowing tocomparatively less valuein RMSE and MAPE measures(Table 3). The P value of theLjung-Box test was greater than 1 %, might accept the null hypothesis that the white noise of ARIMA_GA models' residuals (Table 4). The actual and fitted values of best fitted model of chilli and turmeric in India is shown in Fig. 5-6.

Table3. Comparing the 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 4. 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 explains the probabilistic or stochastic nature of the observation over period of particular time(economic time series). The probability value in ADF test of actual chilli and turmeric productions was higher than 0.05 which supporting to accept null hypothesis as non-stationarity. But the P value of differenced series has less than 0.01 indicating the stationarity [7]. The fitted ARIMA model for chilli production, ARIMA (2,1,1), differs from the findings of Padmanaban et al.[12], whose model ARIMA (0,1,1) was identified as the best for chilli production in India within the period of 1970-2012. This variance can be attributed to the different durations of data analysed in the respective studies.Though ARIMA model has capture the past linear relationship effectively in the system and their parameters are estimated by maximum likelihood estimation, the

Comment [AA37]: Check for the correct table and figure no in the body of the work and Table or figure itself.

Comment [AA38]: This statement should also be in the methodology while explained your training and testing data procedure

Comment [AA39]: The title might be rewritten as:
Comparison of **the predictive ability of fitted models**

Comment [AA40]: Check for grammatical error

estimated parameters are unstable and non-significant when the data has outliers or leverage pointsmoreover stuck with principles of parsimony. Therefore, Rathod et al.,[14] applied stochastic global search algorithm 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 result can also be obtained for chilli and turmeric production.The residuals of fitted ARIMA_GA models for both chilli and turmeric production has efficiently supported the Ljung-Box test's null hypothesis which would ensure the reliability of fitted model for forecasting future values, the result is consistent with Abbasi et al.,[1].

Comment [AA41]: Use another grammar

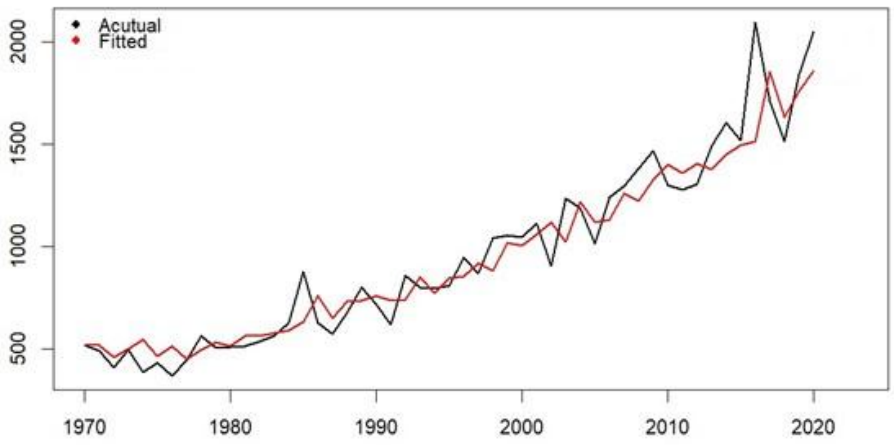

Fig. 5Actual vs fitted value and forecasted of chilli production in India

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, stationarity of the time series was checked by ADF test and the parameters of 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.

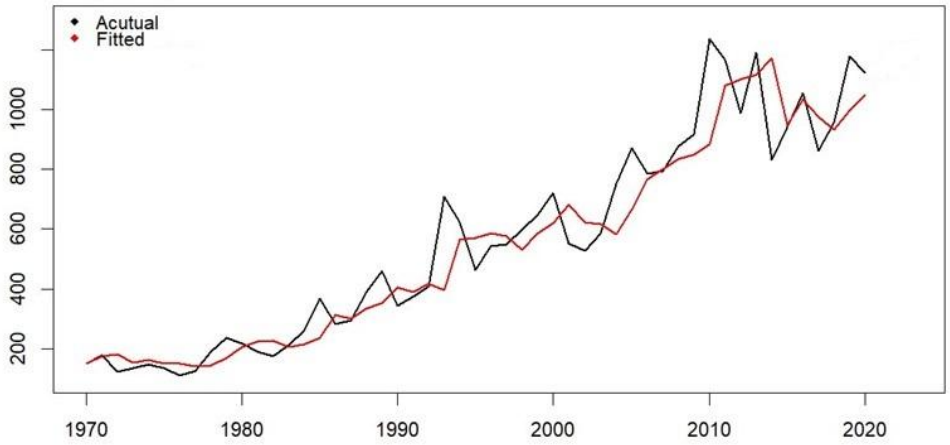

Fig. 6: Actual vs fitted value and forecasted of turmeric production in India

Comment [AA42]: Take it to before conclusion

REFERENCE

1. 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.
2. Alquraish M, Abuhasel K, Alqahtani S, Khadr M. (2021). SPI-based hybrid hidden Markov-GA, ARIMA-GA, and ARIMA-GA-ANN models for meteorological drought forecasting: Sustainability. 2021;13.
3. 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.
4. Box GEP, Jenkins GM. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day; 1970.
5. Brockwell PJ, Davis RA. (1996). Introduction to time series and forecasting. Springer-Verlag: New York; 1996. pp.43-75.
6. Dash A, Mahapatra SK. (2020). 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.
7. DheerP. (2019). Time series modelling for forecasting of food grain production and productivity of India: Journal of Pharmacognosy and Phytochemistry.2019; 8(3): 476-482.
8. Directorate General of Commercial Intelligence and Statistics (DGCI&S). (2020). Available:https://www.indianspices.com/sites/default/files/Major_item_wise_Export_2020.pdf.
9. Gunasekaran V, KoviKK, ArjaS,Chimata R. Solar irradiation forecasting using genetic algorithms: ArXiv preprint *arXiv:2106.13956*.2021; Available: 10.48550/arXiv.2106.13956.
10. Hamjah MA. (2014). Forecasting major fruit crops productions in Bangladesh using Box-Jenkins ARIMA model:Journal of Economics and Sustainable Development. 2014; 5(7): 96-107.
11. Holland J. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: The University of Michigan Press; 1975.
12. Padmanaban K, Sahu PK, Narsimhaiah L. (2016). Production performance of chilli in India- a statistical approach: Advances in Life Sciences.2016; 5(10): 4191-4200.
13. 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.
14. Rathod S, Singh KN, Arya P, RayM, 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. Rolf S, Pravez,J. UrferW. Model identification and parameter estimation of ARMA models by means of evolutionary algorithms:Computational Intelligence for Financial Engineering. 1997; 23: 237-243.

16. Zaer SA, AlsmadiMK,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.

Comment [AA43]: Check the normal reference method