Current Journal of Applied Science and Technology

40(23): 49-57, 2021; Article no.CJAST.74221
ISSN: 2457-1024
(Past name: British Journal of Applied Science & Technology, Past ISSN: 2231-0843,
NLM ID: 101664541)

Hybrid Time Series Models for Forecasting Maize Production in India

Pramit Pandit[1], Bishvajit Bakshi[2*] and Varun Gangadhar[3]

[1]Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, West Bengal, India.
[2]Centre for Management of Health Services, Indian Institute of Management-Ahmedabad, Gujarat, India.
[3]Department of Agricultural Statistics, Applied Mathematics and Computer Science, University of Agricultural Sciences, Bengaluru, Karnataka, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/CJAST/2021/v40i2331492
Editor(s):
(1) Dr. Orlando Manuel da Costa Gomes, Lisbon Accounting and Business School (ISCAL), Lisbon Polytechnic Institute, Portugal.
(2) Dr. Ahmed Mohamed El-Waziry, Alexandria University, Egypt and King Saud University, Saudi Arabia.
Reviewers:
(1) Parag Shah, India.
(2) Abdullah-Al-Zabir, Bangladesh.
Complete Peer review History: https://www.sdiarticle4.com/review-history/74221

Original Research Article

Received 04 July 2021
Accepted 14 September 2021
Published 15 September 2021

ABSTRACT

In spite of the immense success of different linear and non-linear time series models in their respective domains, real-world data are rarely pure linear or non-linear in nature. Hence, a hybrid modelling framework with the capability of handling both linear and non-linear patterns can substantially improve the forecasting accuracy. With this backdrop, an effort has been made in this investigation to evaluate the suitability of hybrid models in compassion to single linear or non-linear models for forecasting maize production in India. Data from 1949-50 to 2016-17 have been utilised for the model building purpose while retaining the data from 2017-18 to 2019-20 for the post-sample accuracy assessment. Outcomes emanated from this investigation clearly reveals that the ARIMA-NLSVR model has outperformed all other candidate models employed in this study. It is noteworthy to mention that both the hybrid models have performed better than their individual counterparts. The superior forecasting ability of both the non-linear models over the linear ARIMA model has also been evident.

____

[*]Corresponding author: E-mail: bishvajitb93@gmail.com;

Keywords: Agricultural forecasting; ARIMA; Hybrid models; Maize production; NLSVR; TDNN.

1. INTRODUCTION

Accurate and reliable forecasting of food grain production is of prime importance for ensuring the food security of the ever-increasing global populace. According to the FAO [1], global food production would be needed to be increased by more than 70% by 2050. Next to only rice and wheat, maize (*Zea mays*) is the most consumed staple food crop in the world [2]. It is considered as the queen of cereals and has utility in many industrial applications [3].

Despite the fact that India has achieved a decent level of food grain production, the gains of production have largely been negated by its population expansion. Earlier maize was mostly cultivated for food consumption purposes, however, due to the changes in Indian dietary patterns, it is now largely being grown for feed purposes [4]. Based on the increasing growth rate of poultry, livestock, fish, and milling industries, the demand for maize is expected to be 45 million tonnes by 2030 [5]. Maize is produced throughout the year in India in a variety of environments ranging from extreme semi-arid to sub-humid and humid conditions. It is also very popular in the low and mid-hill areas of the western and north-eastern regions. Maize growing areas in India can be broadly classified into two categories, viz., (i) traditional maize growing areas (Bihar, Madhya Pradesh, Rajasthan, and Uttar Pradesh) and (ii) non-traditional maize growing areas (Karnataka and Andhra Pradesh). Despite a positive compound annual growth rate (3.88% from 1949-50 to 2019-20), the growth in maize production in India is limited by several constraints such as extreme weather conditions, increased pest and disease incidence, imbalanced use of nutrients, limited adoption of improved technologies etc. [6]. However, due to the lack of region-specific long-term data on several production factors, time series models are usually employed for the forecasting purpose to provide an aid for decision making and in-time planning.

One of the most widely used models for time series forecasting is the autoregressive integrated moving average (ARIMA) model [7-9]. Its immense popularity stems from its sound statistical foundation and the well-known Box–Jenkins method [10]. The major drawback of this model is the presumption of linearity, i.e., no non-linear patterns can be recognised by it. To deal with the non-linear patterns, time-delay neural network (TDNN) and non-linear support vector regression (NLSVR) models are the popular choices. Taskaya-Temizel and Casey [11] have conducted a comparative study of autoregressive neural network hybrids using nine monthly time series. Hadipour et al. [12] have forecasted the groundwater Level in Qom Plain, Iran by employing the TDNN models. Fung et al. [13] have utilized improved support vector regression models for agricultural drought prediction at downstream of Langat River Basin, Malaysia. However, the real-world data are rarely pure linear or non-linear in nature; they often contain both linear as well as non-linear patterns in their structure. Hence, a hybrid modelling framework with the capability of handling both linear and non-linear patterns can substantially improve the forecasting accuracy [14]. Consequently, the applications of the hybrid models are gaining momentum day-by-day. Faruk [15] has mentioned the superiority of the hybrid models over the ARIMA and neural network models for water quality predictions. Khairalla and AL-Jallad [16] have also obtained similar results in the case of financial time series analysis. Results obtained by Rathod et al. [17] have revealed that the forecasting accuracy of the hybrid models is better as compared to the single models and among the hybrid models, the ARIMA-NLSVR model has performed superior than the ARIMA-TDNN model.

In spite of the immense success of different time series models in their respective domains, the empirical evaluations have often yielded mixed results. With this backdrop, an effort has been made in this investigation to evaluate the suitability of hybrid models in compassion to single linear or non-linear models for forecasting maize production in India. The rest of the paper proceeds as follows. The next section provides a detailed description of the data set as well as of the time series models under study. The subsequent section presents the empirical results and discussion. The last section, finally, concludes the paper.

2. MATERIALS AND METHODS

2.1 Data and Statistical Software

Yearly data on maize production (in million tonnes) from 1949-50 to 2019-20 have been collected from ‘Agricultural Statistics at a Glance

- 2020' published by the Directorate of Economics and Statistics, Department of Agriculture, Cooperation & Farmers Welfare, Ministry of Agriculture & Farmers Welfare, Government of India. Production data from 1949-50 to 2016-17 have been utilised for the model building purpose while retaining the data from 2017-18 to 2019-20 for the post-sample accuracy assessment.

The statistical software R and Python have been utilised for modelling and forecasting maize production in India.

2.2 Methodology

2.2.1 ARIMA models

In an autoregressive integrated moving average (ARMA) model, the future value of a variable is assumed to be a linear function of several past observations and random errors [10]. An ARIMA model that can represent homogeneous non-stationary behaviour is written as follows:

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

where y_t and ε_t are the actual observation and random error at time period t, respectively; ϕ_i (i = 1, 2, …, p) and θ_j (j = 1, 2, …, q) are model parameters. p and q are integers and often referred to as orders of the model. B is the backshift operator defined by $By_t = y_{t-1}$ and d represents the order of differencing. Random errors ε_t are assumed to be independent and identically distributed with a mean zero and a constant variance σ^2.

The Box–Jenkins methodology (i.e., ARIMA methodology) includes three iterative steps, viz. (i) model identification, (ii) parameter estimation and (iii) diagnostic checking. At the identification stage, based on autocorrelation and partial autocorrelation patterns, one or several potential models are identified. Once a tentative model is specified, model parameters are estimated such that an overall measure of errors is minimised. At the diagnostic checking stage, the white noise test for the residuals of the tentatively identified candidate model is carried out. If residuals are not white noise, again a candidate model is selected and the same procedure is repeated unless a valid model is found.

2.2.2 TDNN models

The human brain can be considered as a highly complex, non-linear, parallel computer, which served as an inspiration for the basic structure of ANN [18]. Like neurons in the brain, ANN is composed of a number of simple but highly interconnected processing elements [19]. Models based on neural network architecture is like a black box consisting of a series of equations for calculating the output by using the provided input values. A general neural network architecture consists of (i) an input layer that accepts external information, (ii) one or more hidden layer that provides non-linearity to the model and (iii) an output layer that provides the target value. Each layer contains one or more nodes. All the layers in a multilayer neural network are connected through an acyclic arc.

Time series data can be modelled using a neural network in two possible ways. The first way is to explicitly represent time in the form of recurrent connections from output nodes to the preceding layer [20]. The second way is to provide the implicit representation of time, whereby a static neural network like multilayer perceptron is bestowed with dynamic properties [21]. A neural network can be made dynamic by embedding either long-term or short-term memory, depending on the retention time, into the structure of a static network. For temporal data processing, some form of short-term memory is required to make the neural network dynamic. One simple way of building short-term memory into the structure of a neural network is through the use of time delay, which can be implemented at the input layer of the neural network.

The general expression for a multilayer feed-forward time-delay neural network is given by:

$$y_{t+1}=g\left(\sum_{j=0}^{q}\alpha_j\, f\left(\sum_{i=0}^{p}\beta_{ij}\; y_{t-i}\right)\right) \qquad (2)$$

where f and g denote the activation function at the hidden and output layer, respectively. p is the number of input nodes (tapped delay), q is the number of hidden nodes, β_{ij} is the weight attached to the connection between the ith input node and the jth hidden node, α_j is the weight attached to the connection from the jth hidden node to the output node and y_{t-i} is the ith input (lag) of the model. Each node of the hidden layer receives the weighted sum of all inputs including a bias term. This weighted sum of input variables is then transformed by each hidden node using the activation function f. In a similar fashion, the output node also receives the weighted sum of the output of all hidden nodes and produces an output by transforming the weighted sum using its activation function g. In time series analysis, f is often chosen as logistic sigmoid function and g as an identity function.

For a univariate time series forecasting problem, past observations of a given variable serve as the input variables. The neural network model attempts to map the following function:

$$y_{t+1} = f\left(y_t, y_{t-1}, \dots, y_{t-p+1}, w\right) + \varepsilon_{t+1} \qquad (3)$$

where y_{t+1} pertains to the observation at time $t+1$, p is the number of lagged observation, w is the vector of network weights and ε_{t+1} is the error term at time $t+1$. Hence, the neural network acts like a non-linear autoregressive model.

2.2.3 NLSVR models

Support vector machine (SVM) was originally developed for the classification problems. With the introduction of Vapnik's ε-insensitive loss function [22], it has further been extended to the domain of non-linear regression estimation problems and the models developed for solving such problems are known as NLSVR models. The basic concept is to transform the original input space into a high dimensional feature space and then construct linear regression in the newly formed feature space, which corresponds to non-linear regression in the original dimensional input space. Let us consider a vector of data set $\mathrm{Z} = \{\mathrm{x_i}\, \mathrm{y_i}\}_{\mathrm{i=1}}^{\mathrm{N}}$, where $\mathrm{x_i} \in \mathrm{R^n}$ is the input vector, $\mathrm{y_i}$ is the scalar output and N is the size of the data set. The general equation of the NLSVR estimation function is given as follows:

$$\mathrm{f(x)} = \mathrm{w^T}\phi(\mathrm{x}) + \mathrm{b} \qquad (4)$$

where ϕ(.) is a non-linear mapping function, which maps the original input space into a higher dimensional feature space vector. w is the weight vector, which is normal to the hyperplane. b is the bias term and T denotes the transpose.

The performance of NLSVR models highly relies on the kernel function and a set of hyper-parameters. The most commonly used kernel function is the radial basis function (RBF), which requires the optimisation of two hyper-parameters, viz., (i) the regularisation parameter C and (ii) the kernel bandwidth parameter γ. The former one balances the model complexity and approximation accuracy, whereas the latter one defines the variance of the RBF kernel function [23].

2.2.4 Hybrid models

The hybrid methodology considers the time series $\mathrm{y_t}$ as a combination of both linear and non-linear components [14].

$$\mathrm{y_t} = \mathrm{L_t} + \mathrm{N_t} \qquad (5)$$

where, $\mathrm{L_t}$ and $\mathrm{N_t}$ represent the linear and non-linear components present in the time series data, respectively. These two components are to be estimated from the data. This hybrid method of combining forecasting has the following steps;

(i) First of all, a linear time series model (say, ARIMA) is fitted to the data.

(ii) At the next step, the residuals are obtained from the fitted linear model. The residuals will now contain only the non-linear components. Let $\mathrm{e_t}$ denotes the residual at the time t from the linear model, then

$$\mathrm{e_t} = \mathrm{y_t} - \widehat{\mathrm{L_t}} \qquad (6)$$

Where, $\widehat{\mathrm{L_t}}$ is the forecast value for the time t obtained from the linear model.

(iii) Diagnosis of residuals is carried out to check if there is still any linear correlation structure left in the residuals. The residuals are then tested for the possible presence of non-linearity by using the BDS test [24].

(iv) Once the residuals confirm non-linearity, the residuals are modelled using a non-linear model (say, TDNN or NLSVR). The forecast values, $\widehat{\mathrm{N_t}}$ are then obtained for the residual series.

(v) Finally, the forecasted linear and non-linear components are combined to obtain the aggregated forecast values as:

$$\widehat{\mathrm{y_t}} = \widehat{\mathrm{L_t}} + \widehat{\mathrm{N_t}} \qquad (7)$$

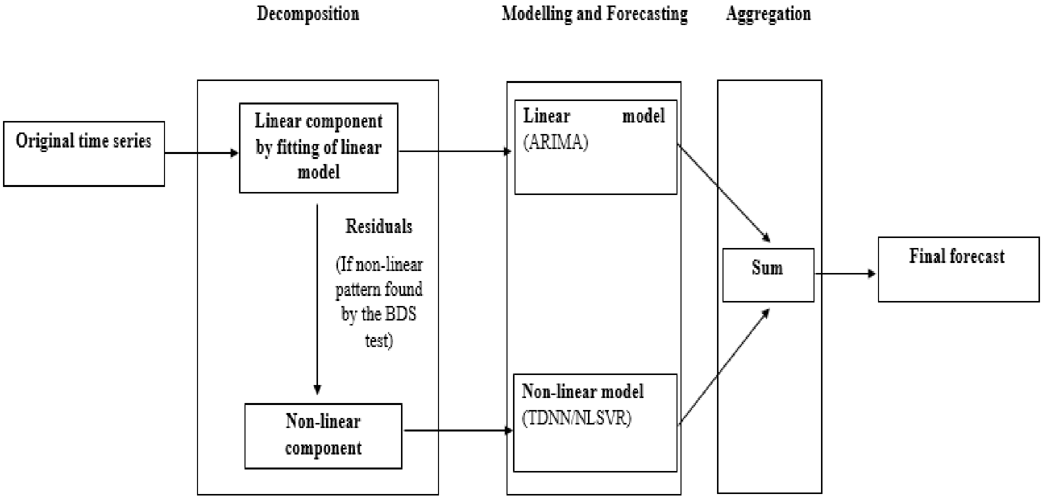
Decomposition
Modelling and Forecasting
Aggregation
Original time series
Linear component by fitting of linear model
Residuals
(If non-linear pattern found by the BDS test)
Non-linear component
Linear model (ARIMA)
Non-linear model (TDNN/NLSVR)
Sum
Final forecast

Fig. 1. Schematic representation of hybrid modelling

2.2.5 Measures of post-sample model accuracy

The forecasting ability of both models is assessed in terms of the two common accuracy measures, viz., root mean squared error (RMSE) and mean absolute percentage error (MAPE). RMSE measures the overall performance of a model and has the form:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \qquad (8)$$

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

$$\mathrm{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \% \qquad (9)$$

3. RESULTS AND DISCUSSION

The first and foremost step in time series analysis is to plot the data. An upward trend over the years is clearly evident from Fig. 2 indicating a possible non-stationarity. Table 1 briefs the descriptive statistics. The average production of maize is observed to be 10.07 million tonnes during the study period. A high coefficient of variation value entails a higher degree of instability in the data. Results of the ADF (Augmented Dickey-Fuller) test, as presented in Table 2, clearly indicate the non-stationary and stationary nature of the level and the first difference series, respectively.

Table 1. Descriptive statistics

Statistics	Value
Mean	10.07
Minimum	1.73
Maximum	28.77
Standard deviation	7.22
Skewness	1.20
Kurtosis	0.46
Coefficient of Variation (%)	71.76

Table 2. Results of the ADF test

Series	Test statistic	p value
Level series	2.128	0.999
First difference series	-4.039	0.013

Among the different candidate ARIMA models, ARIMA (1, 1, 2) model is finally selected on the basis of least AIC (Akaike information criterion) and BIC (Bayesian information criterion) values as well as least RMSE and MAPE values. Due weightage has been given to the well-behaved residuals. Parameter estimates of the selected ARIMA (1, 1, 2) model are provided in Table 3. In the next step, the residuals obtained from the selected ARIMA models are tested for the presence of non-linearity. Results of the BDS test, as provided in Table 4, clearly confirms the presence of non-linear patterns in its structure. Hence, TDNN and NLSVR models are employed for both the original and residual series.

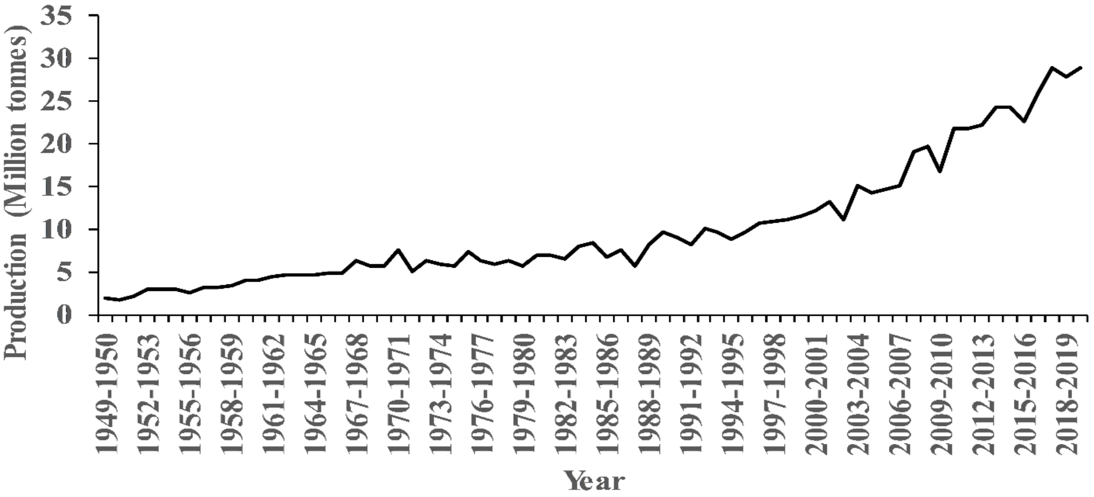

Fig. 2. Maize production (in million tonnes) in India during 1949-50 to 2019-20

Table 3. Parameter estimates of the ARIMA model

Parameters	Estimate	p value
C (Constant)	0.34	0.01
ϕ_1	0.82	<0.01
θ_1	-1.59	<0.01
θ_2	0.79	<0.01

Table 4. Results of the BDS test

Epsilon	Embedding dimension (m=2)		Embedding dimension (m=3)	
	Statistic	p value	Statistic	p value
3.11	23.84	<0.01	36.17	<0.01
6.22	16.23	<0.01	18.69	<0.01
9.34	13.69	<0.01	14.33	<0.01
12.45	11.50	<0.01	11.68	<0.01

In the case of TDNN models, the number of input and hidden nodes have been varied from 1 to 6 and from 2 to 14, respectively. Keeping training and testing accuracy as well as parsimony in mind, the TDNN model with three input nodes and two hidden nodes (3:2s:1l) has been selected for the original (first difference) series. However, for the ARIMA residual series, the TDNN model with two input nodes and seven hidden nodes (2:7s:1l) has performed better than the other candidate TDNN models. Specifications of both these models are furnished in Table 5. Similar to the TDNN models, the NLSVR models for the original and ARIMA residual series have been built with the model specifications presented in Table 6.

Table 5. Specifications of the TDNN models

Specifications	Original series	Residual series
No. of input nodes	3	2
No. of hidden nodes	2	7
Total no. of weights	11	29
Hidden layer activation function	Logistic	Logistic
Output layer activation function	Identity	Identity

Table 6. Specifications of the NLSVR models

Series	Kernel function	No. of SVs	C	γ	ε
Original series	RBF	13	100	0.25	0.01
Residual series	RBF	12	100	0.25	0.01

Table 7. Post-sample assessment of model accuracy

Model	RMSE	MAPE
ARIMA	2.19	7.42
TDNN	1.92	4.75
NLSVR	0.83	2.82
ARIMA-TDNN	1.49	4.19
ARIMA-NLSVR	0.01	0.02

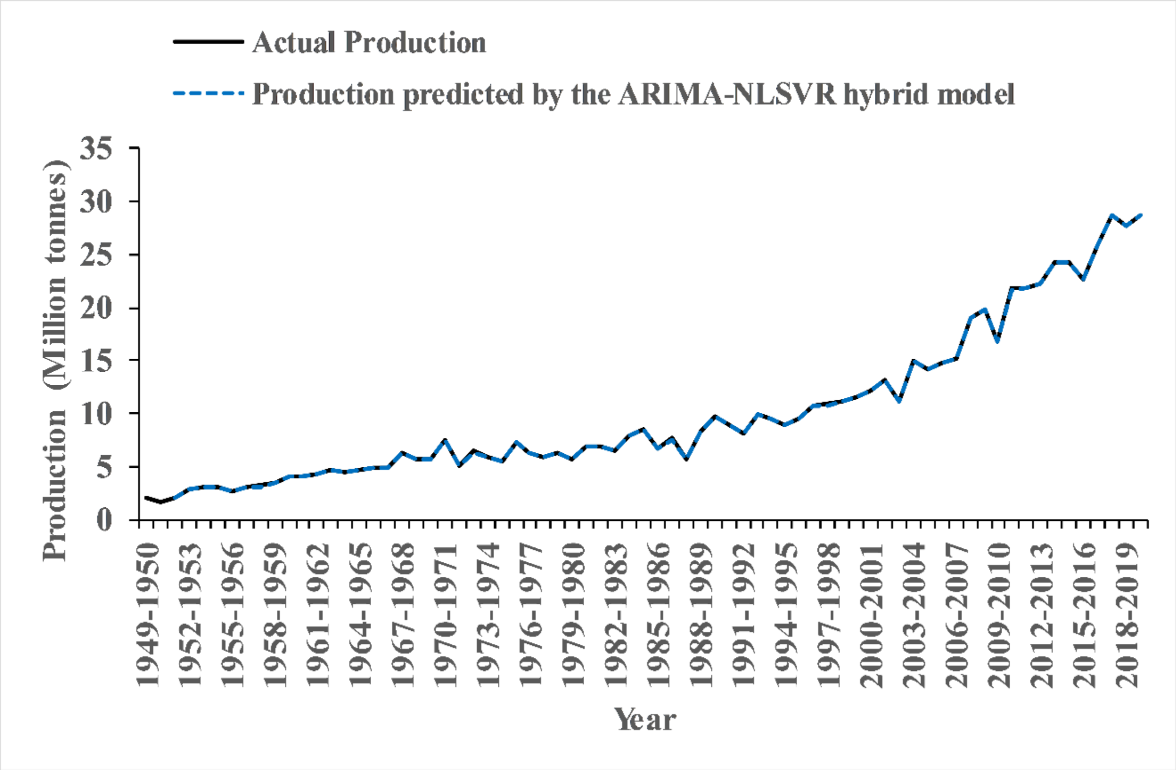

Fig. 3. Actual and predicted (by the ARIMA-NLSVR hybrid model) maize production in India

Comparative assessment of the different models in terms of post-sample RMSE and MAPE values is presented in Table 7. It can be clearly observed that the ARIMA-NLSVR model has outperformed all other candidate models in forecasting maize production in India. The actual and predicted maize production by the best fitted ARIMA-NLSVR hybrid model is presented in Fig. 3. It is noteworthy to mention that both the hybrid models have performed better than their individual counterparts [17, 25-27]. The superior forecasting ability of both the non-linear models over the linear ARIMA model has also been evident.

4. CONCLUSION

The current study has compared the hybrid models with single linear or non-linear models for forecasting maize production in India. Among the competing time series models, the ARIMA-NLSVR hybrid model has performed substantially better than the others. We also find that the non-linearity test has provided fair guidance in the application of the non-linear models. However, the generalisation of this study includes different linear and non-linear models in the hybrid modelling framework.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. FAO. How to Feed the World in 2050; 2009.
   Available:http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf.
2. Nuss ET, Tanumihardjo SA. Maize: A paramount staple crop in the context of global nutrition. Compr Rev Food Sci Food Saf. 2010;9(4):417-36.
3. Kaul J, Jain K, Olakh D. An overview on role of yellow maize in food, feed and nutrition security. Int J Curr Microbiol Appl Sci. 2019;8(2):3037-48.
4. Hellin J, Erenstein O. Maize-poultry value chains in India: implications for research and development. J New Seeds. 2009;10(4):245-63.
5. FICCI. India maize summit'15; 2015.
   Available: https://ficci.in/past-event-page.asp?evid=22310.
6. Kumar R, Srinivas K, Sivaramane N. Assessment of the maize situation, outlook and investment opportunities in India. National Academy of Agricultural Research Management, Hyderabad, India: Country report–regional assessment Asia (MAIZE-CRP); 2013.
7. Sarika, Iquebal, MA, Chattopadhyay C. Modelling and forecasting of pigeon pea (*Cajanus cajan*) production using autoregressive integrated moving average methodology. Ind J Agric Sci. 2011;81(6):520-3.
8. Suresh KK, Priya SRK. Forecasting sugarcane yield of Tamil Nadu using ARIMA models. Sugar Tech. 2011;13(1):23-6.
9. Kumari P, Mishra GC, Pant AK, Shukla G, Kujur SN. Autoregressive Integrated Moving Average (ARIMA) approach for prediction of rice (*Oryza sativa* L.) yield in India. BioScan. 2014;9(3):1063-6.
10. Box GEP, Jenkins G. Time series analysis, forecasting and control. San Francisco, CA: Holden-Day; 1970.
11. Taskaya-Temizel T, Casey MC. A comparative study of autoregressive neural network hybrids. Neural Netw. 2005;18(5-6):781-9.
12. Hadipour A, Khoshand A, Rahimi K, Kamalan HR. Groundwater level forecasting by application of artificial neural network approach: A case study in Qom Plain, Iran. J Hydro-Environ Res. 2019;3(5):30-4.
13. Fung KF, Huang YF, Koo CH, Mirzaei M. Improved SVR machine learning models for agricultural drought prediction at downstream of Langat River Basin, Malaysia. J Water Clim Chang. 2020;11(4):1383-98.
14. Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neural Comput. 2003;50:159-75.
15. Faruk DÖ. A hybrid neural network and ARIMA model for water quality time series prediction. Eng Appl Artif Intell. 2010;23(4):586-94.
16. Khairalla M, AL-Jallad NT. Hybrid forecasting scheme for financial time-series data using neural network and statistical methods. Int J Adv Comput Sci Appl. 2017;8(9):319-27.
17. Rathod S, Mishra GC, Singh KN. Hybrid time series models for forecasting banana production in Karnataka State, India. J Indian Soc Agric Stat. 2017;71(3):193-200.
18. Shanmuganathan S, Samarasinghe S, editors. Artificial neural network modelling. Switzerland: Springer Nature; 2016.
19. Tosun E, Aydin K, Bilgili M. Comparison of linear regression and artificial neural network model of a diesel engine fueled with biodiesel-alcohol mixtures. Alex Eng J. 2016;55(4):3081-9.
20. Elman JL. Finding structure in time. Cog Sci. 1990;14:179-211.
21. Haykin S. Neural networks – a comprehensive foundation. Upper Saddle River: Prentice-Hall; 1999.
22. Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation, and signal processing. In: Mozer M, Jordan M, Petsche T, editors. Advances in neural information processing systems. Cambridge: MIT Press; 1997.
23. Vapnik V. The nature of statistical learning theory. 2nd Edition. New York: Springer-Verlag; 2000.
24. Brock WA, Dechert WD, Scheinkman JA, lebaron B. A test for independence based on the correlation dimension. Econom Rev. 1996;15:197-235.
25. Xu D, Zhang Q, Ding Y, Huang H. Application of a Hybrid ARIMA–SVR Model

Based on the SPI for the Forecast of Drought—A Case Study in Henan Province, China. J Appl Meteorol Climatol. 2020;59(7):1239-59.

26. Pannakkong W, Huynh VN, Sriboonchitta S. A novel hybrid autoregressive integrated moving average and artificial neural network model for cassava export forecasting. Int J Comput Intell Syst. 2019;12(2):1047-61.

27. Shin JY, Kim KR, Ha JC. Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management. Agric For Meteorol. 2020;281:107858-73.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle4.com/review-history/74221