Ann. For. Res. 65(1): 9119-9133, 2022
ISSN: 18448135, 20652445

ANNALS OF FOREST RESEARCH
https://www.e-afr.org/

DEVELOPMENT OF HYBRID TIME SERIES MODELS FOR FORECASTING AUTUMN RICE USING ARIMAX- ANN AND ARIMAX-SVM.

Borsha Neog*, Bipin Gogoi[1], A.N. Patowary[2]
*Assistant Professor, Department of Agril. Statistics, Assam Agricultural University, Jorhat-13
[1] Professor, Department of statistics, Dibrugarh University, Dibrugarh-786004
[2]Assistant Professor, College of Fishery Science, Assam Agricultural University, Raha-782103
Email: borsha.neog@aau.ac.in

Abstract:

Time series forecasting is a very active research topic in the domain of science and engineering. The study of forecasting in time series analysis has become a powerful tool in different applications in the agricultural field. Climate change is another major concern in world wide, many researchers are trying to understand its impact on growth and production of crops. Keeping this importance of climate on development and production of crops, an attempt has been made to develop the time series models for forecasting production of Autumn rice in Assam. Yearly data on production of all selected crop and all selected climatic variables have been used for forecasting from the year 1981 to 2018. The data from 1981-2007 were used for model building and 2008 - 2018 were used for checking the forecasting performance of the model. The statistical software viz., SPSS, R; were used for modelling and forecasting of production of agricultural crops in Assam. In this study ARIMAX, ANN, SVM time series models and hybrid of both ARIMAX-ANN, ARIMAX-SVM were used to analyse the past behaviour of production of Autumn rice related with selected climatic variables in order to make inferences about its future behaviour. ARIMA (2,1,2) for production of Autumn rice is applied along with all the weather variables over the growth period of the crop for estimation of production of Autumn rice. The value of MAE under training set for different models ARIMAX (2,1,2), ANN (03:4s:1l), SVM, ARIMAX-ANN and ARIMAX-SVM are found to be 38705.376, 35438.910, 34335.563, 31967.161 & 29484.631 respectively, whereas the value of MAE under testing set are found to be 13502.977, 12430.576, 11149.712, 9436.347 & 9176.531 respectively. Based on these results the model ARIMAX-SVM can be recommended for forecasting of production of crop because of the minimum value of MAE both under training and testing set.

Key Words: Autumn Rice, Forecast, Time series, Hybrid.

1. Introduction:

Forecasts can be formed in various ways. Different statistical approaches like regression, time-series and stochastic models are in vogue for forecasts of any crop. Every statistical approach has its own advantages and limitations. Regression analysis is the most widely used statistical methods for modelling the relationship between variables. Some of the applications of regression analysis involves regressor and response variables that have an order of sequence over time and then the necessity of time series modelling arises for the analysis of such order of sequence. Time series models have some advantages in certain situations. They can be used easily for forecasting

purposes because of availability of past data with equally spaced intervals over discrete point of time. These successive observations are statistically dependent and time series modelling is concerned with different techniques for the analysis of such dependence. In the field of agriculture, forecasting of different variables for different crops or regions have immense importance. Several researchers have developed different forecast models based on time series data using different methodologies such as time series decomposition models, exponential smoothing models, seasonal ARIMA models, VARX models, vector ARMA models using multivariate time series etc. When an ARIMA model includes other time series as input variables, referred as ARIMAX model, the response series is modelled using the current and past values of exogenous variable(s) as input series. Padhan (2012) applied ARIMA models for forecasting the productivity of 34 selected agricultural products of India from the year 1950 to 2010.

Climate change is another major concern in worldwide, many researchers are trying to understand its impact on growth and production of crops, and identifying the suitable management options to sustain the production of crops under estimated climate change scenario. Quantitative understanding of crop response to climate needs the development of different statistical models for various characteristics of the crop by considering its time sequence behaviour along with exogenous climatic factors. Chadsuthi et al. (2012) have applied multivariate ARIMA (ARIMAX) model and showed that the factoring in rainfall with 8 months lag yielded the best model for the northern region while the model factoring in rainfall with 10 months lag and temperature with 8 months lag yielded best model for the north-eastern region. The major problem is how to incorporate the pertinent external information into the forecasting process and subsequently into the decision-making process. But due to presumption of linearity we cannot take this model for nonlinear time series. In real, time series data are nonlinear in nature, so for prediction of nonlinear data, neural network is most promising and potential techniques. Zhang's first paper published on combination of both models, which is known as Hybrid method (Zhang 2003). The Hybrid Approach assumes that linear and non-linear components are additive in nature. Moreover, combination of Neural Network with ARIMA model often yield superior results in case of forecasting performance (Jha and Sinha 2014, Rathod et al. 2017, Ray et al. 2016, Ayub and Jafri 2020).

Keeping this importance of climate on development and production of crops, an attempt has been made to develop the time series models for forecasting production of Autumn rice in Assam.

2. Data and methodology:

To obtain the results we have used Autoregressive integrated moving average using exogenous variable (ARIMAX) model for forecasting of production using weather variables. ARIMAX is an acronym for autoregressive integrated moving average with exogenous variables. It is a logical extension of pure ARIMA modelling with additional variable or exogenous variable. simply, it is a merging of regression and ARIMA. When the term AR and MA are not sufficient to provide an overall explanatory power of an ARIMA model then the model includes different time series as input variables referred as an ARIMAX model fill-up the gap of the model. The forecast values of

production through ARIMAX model were improved by hybrid approaches to compute forecast of production up to the desired year.

Here an attempt has been taken to develop more efficient hybrid model ARIMAX through neural network and support vector machine for agricultural crop production. Yearly data on production of all selected crop and all selected climatic variables have been used for forecasting from the year 1981 to 2018. The data from 1981-2007 were used for model building and 2008 - 2018 were used for checking the forecasting performance of the model. The statistical software viz., SPSS, R; were used for modelling and forecasting of production of agricultural crops in Assam. SPSS software was used to build the suitable ARIMAX model nonlinearity test for residuals obtained from ARIMAX models using Ljung-Box test. R software package 'Forecast' was used for modelling and forecasting using NN and package 'e 1071' was used for modelling and forecasting using SVM.

To develop an ARMAX model, the first step is consisting of identifying a suitable ARMA model for the endogenous variable. Stationarity test of exogenous variable requires before modelling the ARMAX model. Here Nonlinear least square estimation procedure is applied to estimate the parameters of ARMAX model. In the model ARMAX, production of agricultural crop was considered as dependent variable while all the weather variables as exogenous variables. To this end, forecast of covariates using hybrid time domain approaches and neural network are utilized the ARIMAX model.

2.1 Time series forecasting models:

2.1.1 The ARIMAX model:

ARIMAX is an acronym for autoregressive integrated moving average with exogenous variables. It is a logical extension of pure ARIMA modeling that incorporates independent variables which add explanatory value. Conceptually, it is a merging of regression and ARIMA modeling.

When both the terms AR & MA in a pure ARIMA model are not sufficient to provide an acceptably overall explanatory power of a model, it is only natural to look for other driving phenomena whose influence over time is not sufficiently embedded in the historical values of the dependent time series. When an ARIMA model adds other time series as input variables, the model is sometimes referred to as an ARIMAX model. i.e., in addition to past values of the response series and past errors, the response series is modelled using the current and past values of input series.

An ARIMAX form of the model is presented as:

$\phi(B)\, y_t = \beta x_t + \theta(B)\, a_t$ or $y_t = \frac{\beta}{\phi(B)}\, x_t + \frac{\theta(B)}{\phi(B)} a_t$

where x_t is a covariate at time t and β is its coefficient. β can only be interpreted conditional on the previous values of the response variable.

$\phi(B) = 1- \phi_1 B-\ldots-\phi_p B^p$ and $\theta(B)= 1-\theta_1 B-\ldots- \theta_q B^q$

For ARIMA errors in case of non-stationary data, $\phi(B)$ is simply replaced with $\nabla^d\phi(B)$. where $\nabla = 1- B$ denotes the differencing operator.

2.1.2 Artificial Neural Network (ANN) model:

ANN(s) models are set of nonlinear models that can capture different nonlinear structures present in the data set. The specification of ANN model does not require any prior assumption of the data generating process, instead it is largely depended on characteristics of the data known as data-driven approach. Single hidden layer feed forward network is the most widely used model for time series modelling and forecasting. This model is constructed by a network of three layers of simple processing units, and thus termed as multilayer ANNs. The first layer is input layer, the middle layer is the hidden layer and the last layer is output layer.

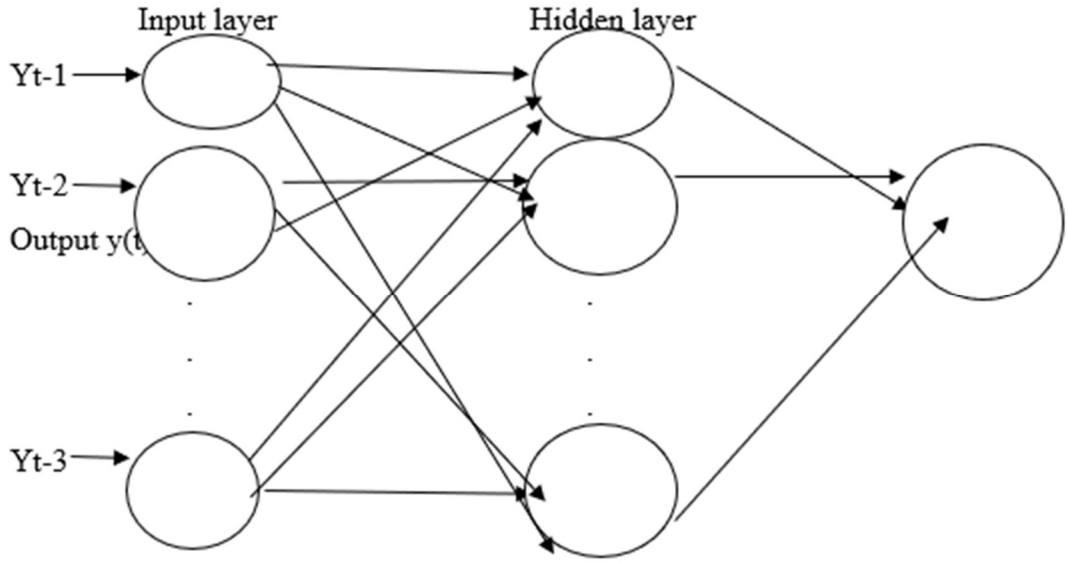

Figure 2.1: Neural Network architecture

The relationship between the output (yt) and the inputs (yt-1,yt-2, …, yt-p) can be mathematically represented as follows:

$Yt= f\left(\sum_{j=0}^{q} w_j g\left(\sum_{i==0}^{q} w_{ij}\ y_{t-i}\right)\right)$ (1)

Where wj(j=0,1,2,…,q) and wij(i=0,1,2,…,p; j=0,1,2,…,q) are the model parameters often called the connection weights; p is the number of input nodes and q is the number of hidden nodes, g and f denote the activation function at hidden and output layer respectively.

2.1.3. Support Vector Machine:

Support vector machine proposed by Vapnik (1998) is a nonlinear algorithm used in supervised learning framework for data classification, pattern recognition and regression analysis. The model has been built in two steps: i.e., training and testing. In the training and testing. In the training step, the largest part of the dataset has been used for the estimation of the function. In the testing step,

the generalization ability of the model has been evaluated by checking the model performance in the small subset.

It has been used in a wide range of applications such as in data mining, classification, regression, and time series forecasting (Cao and Tay, 2001; Flake and Lawrence, 2002; Zhao et al. 2006). The ability of SVM is to solve nonlinear regression estimation problems and it makes SVM successful in time series forecasting.

The SVM architecture is shown in Fig 1.

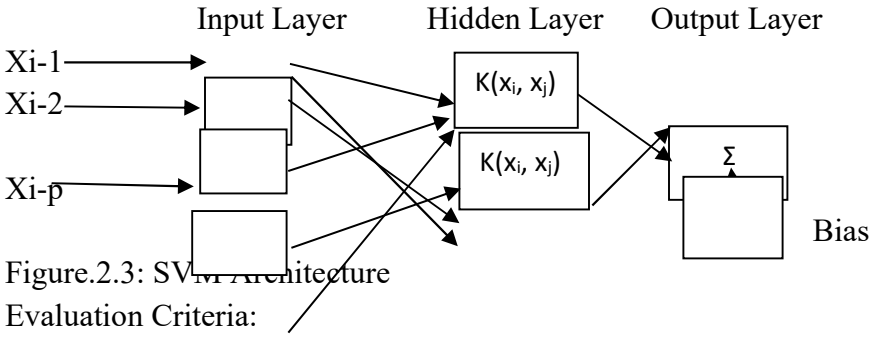

Figure.2.3: SVM Architecture

Evaluation Criteria:

The most common error function in neural networks is the sum of squared errors. Other error functions offered by different software include least absolute deviations, least fourth powers, asymmetric least squares, and percentage differences.

2.1.4. Hybrid approach:

The proposed approach considered time series (yt) as a function of linear and non-linear components. Hence yt= f (Lt, Nt)

where yt is a time series data; Lt and Nt represents the linear and nonlinear component respectively. This approach follows the Zhang's (2003) hybrid approach, accordingly the relationship between linear and nonlinear components can be written as following

Yt= Lt + Nt

The main strategy of this approach is to model the linear and nonlinear components separately by different model. The methodology consists of three steps. Firstly, ARIMA model is applied to the data series to fit the linear part. Let the prediction series provided by ARIMA model denoted as $\widehat{L_t}$. In the second step, instead of predicting the linear component, the residuals denoted as et which are nonlinear in nature are predicted. The residuals can be obtained by subtracting the predicted value $\widehat{L_t}$ from actual value of the considered time series yt.

$$et= yt- \widehat{L_t}$$

Now the residuals are predicted employing an ANN and SVM model. Let the prediction series provided by ANN/SVM model denoted as $\widehat{N_t}$. Finally, the predicted linear and nonlinear components are combined to generate aggregate prediction.

$$\widehat{y_t}= \widehat{L_t} + \widehat{N_t}$$

Ljung-Box test is used to test for non-linearity in this study.

The graphical representation of proposed approach is expressed in the figure 2.2 & 2.3

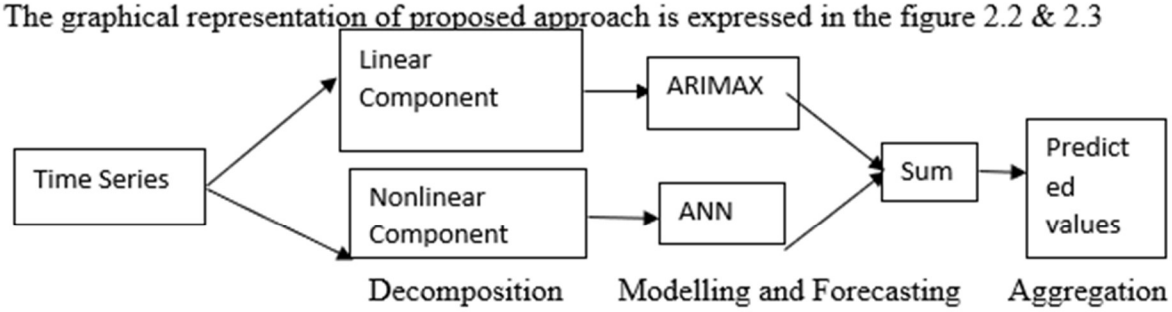

Figure 2.2: Schematic representation of ARIMAX-ANN hybrid methodology

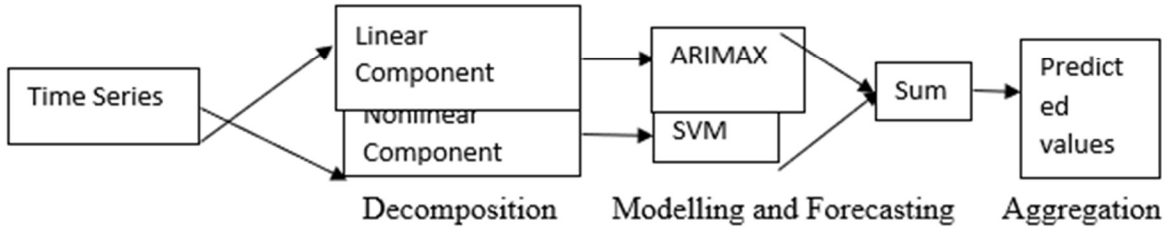

Figure 2.3: Schematic representation of ARIMAX-SVM hybrid methodology

Forecasting Performance:

Forecasting Performance of the model has been adjusted by computing mean absolute error (MAE). The model with minimum values of MAE for training and testing data set is preferred for forecasting purpose. The MAE is computed as

$$\text{MAE}=\frac{1}{n}\sum_{t=1}^{n}|y_t-\hat{y}_t|$$

Where n is the total number of forecast values. Yt is the actual value at period t and $\hat{y}_t$ is the corresponding forecast value.

3. Results and Discussion:

3.3.1. Autumn rice:

The ARIMA models with weather variables as independent variables were applied for fitting ARIMAX models. ARIMA (2,1,2) for production of Autumn rice is applied along with all the weather variables over the growth period of the crop for estimation of production of Autumn rice. Based on stepwise regression, we have selected most influencing weather variables with production of Autumn rice.

Log likelihood test, Akaike Information Criterion (AIC), Schwartz's Bayesian Criterion (SBC), and residual variance were used to estimate the coefficients of AR and MA model. The residual ACF and PACF with t tests and chi squared test suggested by Ljung and Box were applied to check the random shocks to be white noise. The results pertaining to ARIMAX models for production of Autumn rice with different weather variables are presented below:

Table 3.1: Goodness of fit Statistics of Autumn rice

Fit Statistic	ARIMA (2,1,2) with Min. Temp	ARIMA (2,1,2) with Min. Temp and Max. Temp	ARIMA (2,1,2) with Min. Temp, Max. Temp and Precipitation	ARIMA (2,1,2) with Min. Temp, Max. Temp, Precipitation & Windspeed
Stationary R-squared	0.445	0.450	0.449	0.509
R-squared	0.832	0.833	0.833	0.851
RMSE	51569.396	52190.043	53135.451	51036.390
MAPE	8.825	8.636	8.594	8.405
MAE	35347.849	34976.651	34913.06	33978.727
MaxAPE	35.452	32.973	33.154	39.443
MaxAE	121937.074	128041.352	128649.57	112935.561
Normalized BIC	22.287	22.408	22.542	22.559

Table 3.2: Test for white noise of Autumn rice

Model	Ljung-Box Q		
	Statistics	DF	Sig.
ARIMA (2,1,2) with MinTemp	17.823	14	0.215
ARIMA (2,1,2) with MinTemp and MaxTemp	17.353	14	0.238
ARIMA (2,1,2) with MinTemp, MaxTemp and Precipitation	17.255	14	0.243
ARIMA (2,1,2) with MinTemp, MaxTemp, Precipitation & Windspeed	22.608	14	0.067

Table 3.3: Parameter Estimates of Autumn rice

Models				Parameter Estimate	SE	t	Sig.
		Constant		1253698.957	372907.206	3.362	0.002
		AR	Lag1	-1.373	0.203	-6.754	0.000

ARIMA (2,1,2) with Min. Temp, Max. Temp, Precipitation & Windspeed	Production of Autumn rice		Lag2	-0.614	0.210	-2.918	0.007
		MA	Lag1	-0.862	3.453	-.250	0.805
			Lag2	0.135	0.675	.200	0.843
	Min. Temp	Numerator	Lag0	-.46256.485	13883.198	-3.332	0.002
	Max. Temp	Numerator	Lag0	-13690.401	8693.098	-1.575	0.127
	Precipitation	Numerator	Lag0	-186.374	221.278	-0.842	0.407
	Windspeed	Numerator	Lag0	225388.417	102227.827	2.205	0.036

Figure 3.1: Residual ACF and PACF of Autumn rice

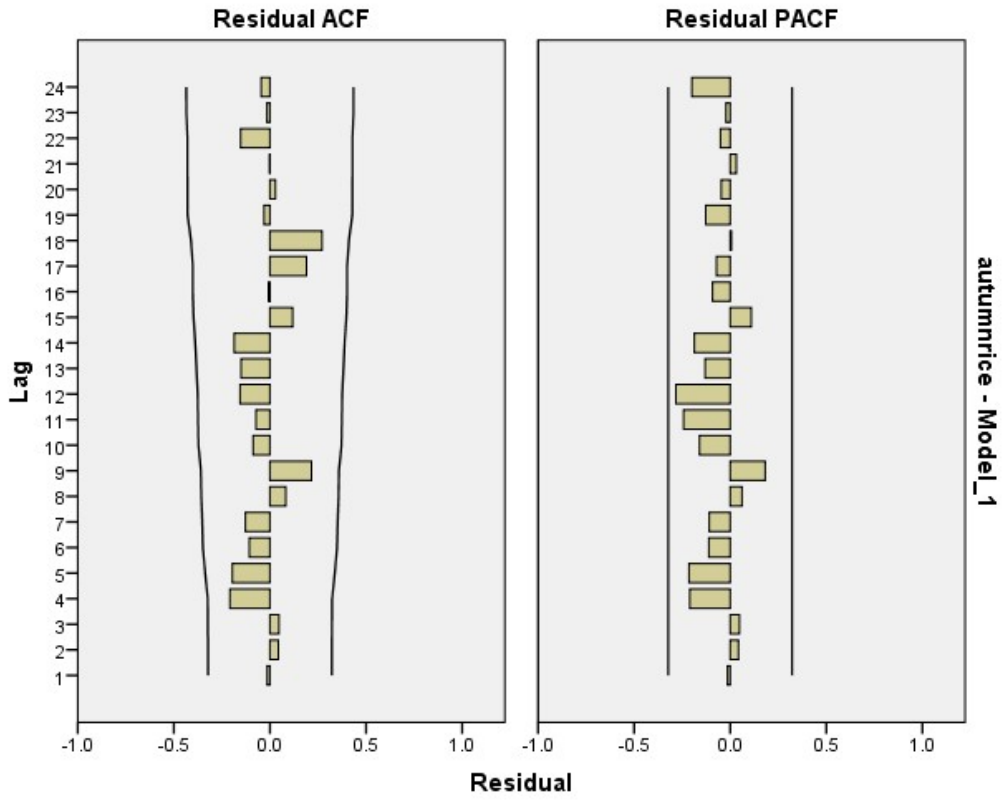

Figure 3.2.: Graphical representation of Forecast of production of Autumn rice

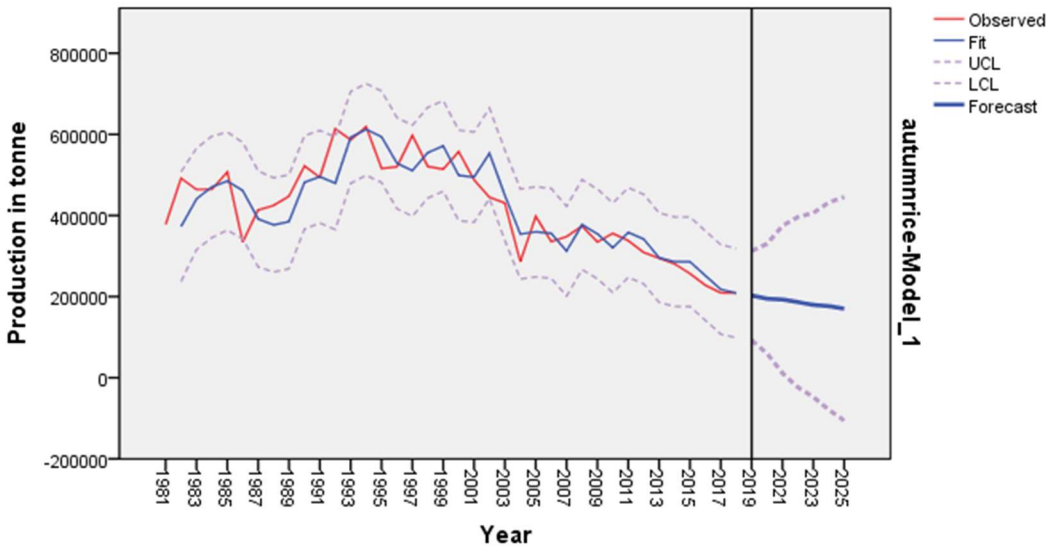

Table 3.4: Forecast of production of Autumn rice

Year	Forecast	LCL	UCL
2019	202498	92334	312662
2020	194805	59519	330090
2021	192860	11281	374439
2022	186233	-23775	396241
2023	179494	-47561	406549
2024	176632	-79032	432296
2025	170121	-105444	445686

Table 3.5: MAE for Neural Network models for production of Autumn rice

Model parameters	MAE for Training	MAE for Testing
1:2s:1l	38454.716	18934.954
1:4s:1l	38450.189	18755.183

1:6s:11	38460.074	18798.951
1:8s:11	38466.646	18851.309
1:10s:11	38468.194	18884.013
1:12s:11	38464.065	18746.649
2:2s:11	37954.681	17346.746
2:4s:11	38046.431	17299.649
2:6s:11	38074.695	17359.603
2:8s:11	38004.383	17345.227
2:10s:11	38039.341	17199.921
2:12s:11	38024.207	17069.593
3:2s:11	39084.747	15341.024
3:4s:11	38705.376	13502.977
3:6s:11	39273.838	13474.494
3:8s:11	39062.022	14224.495
3:10s:11	39145.841	13436.604
3:12s:11	39182.053	13344.056

From the above table, the model 3:4s:11 was found to be the best one based on minimum values of MAE for training= 38705.376 and testing= 13502.977. From this selected model we have got the estimated values of residuals and fitted values of production of autumn rice obtained by ARIMA (2,1,2) with weather variables maximum temperature, minimum temperature, precipitation and windspeed then forecast value of production was obtained through hybrid approach i.e., ARIMAX (2,1,2)-ANN. The goodness of fit measure MAE for hybrid ARIMAX-ANN was found to be 31114.963 as compare to 33978.727 ARIMAX (2,1,2). Residuals obtained by using ARIMAX (2,1,2) were applied on the non-linear approach support vector machine using radial basis function as kernel. Forecast values of production obtain through ARIMAX (2,1,2) were corrected by using the residuals through SVM and estimated the value MAE for hybrid ARIMAX-SVM. MAE for

hybrid ARIMAX-SVM was found to be 28463.427 as compare to 33978.727 of ARIMAX (2,1,2) and 31114.963 of hybrid ARIMAX-ANN. Hence the performance of hybrid model found to be better than ARIMAX (2,1,2) alone.

For the purpose of forecast value of production through hybrid approach, we have got forecast of residuals through the best neural model (03:4s:1l) till 2025. Based on the forecasted value of residuals we found the forecast value of production through hybrid approaches and presented in Table 5.6 along with forecast values by ARIMAX (2,1,2).

Table 3.6: Experimental Results of forecast of Production of Autumn rice

Year	Actual values of Production	Forecast Production by ARIMAX (2,1,2)	Forecast Production by Hybrid Approach using ANN
1981	377857		
1982	491723	415136	
1983	464149	513601	
1984	464585	491791	
1985	507490	467386	468126.6
1986	334881	446113	446015.8
1987	413865	439841	439446.7
1988	424719	405024	401981.4
1989	447598	427678	425315.4
1990	522189	500531	502523.4
1991	494223	460099	463268.6
1992	613696	533757	537265.8
1993	586620	589185	594108.5
1994	619126	579777	584876.2
1995	516032	572177	576315.7

1996	520191	499035	500634
1997	597478	572840	573471.3
1998	520605	520927	522311.4
1999	514156	527136	530034.8
2000	557764	476143	477866
2001	487719	522528	525291.8
2002	444884	479933	483381.3
2003	430474	487270	488563.7
2004	286328	399264	396682.7
2005	398077	372049	366836.1
2006	335708	347161	342480.8
2007	347992	340850	339986.3
2008	374010	367638	369421.4
2009	334655	287485	289029.1
2010	355825	328491	331408.5
2011	338015	315962	319933.1
2012	308745	372859	376948.6
2013	294440	290686	292176
2014	280693	267250	266618.2
2015	256729	232376	232540.8
2016	228146	224148	226743.1
2017	209349	199611	202406.6

2018	209122	215126	217537.3
2019		194805	196562.5
2020		192860	194313.2
2021		186233	187583.1
2022		179494	181032.3
2023		176632	178162.7
2024		170121	171656.4
2025		372551	374091.1

Table 3.7: MAE of different models for production of Autumn rice

Data	ARIMAX	ANN	SVM	ARIMAX-ANN	ARIMAX-SVM
Training	40267.038	38705.376	34335.563	31967.161	29484.631
Testing	12438.631	13502.977	11149.712	9436.347	9176.531

From the above table, the value of MAE under training set for different models ARIMAX (2,1,2), ANN (03:4s:1l), SVM, ARIMAX-ANN and ARIMAX-SVM are found to be 38705.376, 35438.910, 34335.563, 31967.161 & 29484.631 respectively, whereas the value of MAE under testing set are found to be 13502.977, 12430.576, 11149.712, 9436.347 & 9176.531 respectively. Based on these results the model ARIMAX-SVM can be recommended for forecasting of production of crop because of the minimum value of MAE both under training and testing set.

4. Conclusion:

We have applied ARIMAX model on production of Autumn rice along with climatic factors viz., rainfall, maximum temperature, minimum temperature, relative humidity, precipitation, and wind speed as exogenous variables. Based on minimum value of goodness of fit, and stepwise regression; ARIMAX (2,1,2) model with rainfall for autumn rice was found to be the suitable one. P values of parameters estimates of ARIMAX (2,1,2) with rainfall as exogenous variable are estimated to be <0.001 and respectively. Residuals were found to be white noise. Hence, the ARIMAX (2,1,2) model was found to be suitable model under rainfall for Autumn rice. MAE under ARIMAX (2,1,2) with rainfall was estimated to be 33978.727. ANN and SVM approach

were applied on the residuals of ARIMAX (2,1,2) for modelling and forecasting of the residuals. ANN model with (3:4s:1l) was identified as suitable model as this model having minimum values of MAE i.e., 38705.376 & 13502.977 under training and testing data sets respectively. Using 3:4s:1l model, we have estimated the fitted values of residuals and these fitted residuals were used to correct the fitted values of production obtained through ARIMAX (2,1,2) model and eventually get the fitted values under hybrid ARIMAX (2,1,2) model. The MAE under the hybrid ARIMAX-ANN is estimated to be 31114.963 and MAE under ARIMAX-SVM is estimated to be 28463.427. Hence, hybrid ARIMAX-SVM gives better results as compared with hybrid ARIMAX-ANN, ARIMAX. Based on these encouraging results, hybrid ARIMA and ARIMAX using machine learning techniques can be recommended for forecasting of crop production as it has caused significant reduction in MAE for both under training as well as testing sets of data.

References:

1. Anggraeni, W. et al. (2019). Forecasting the Price of Indonesia's Rice Using Hybrid Artificial Neural Network and Autoregressive Integrated Moving Average (Hybrid NNs-ARIMAX) with Exogenous Variables. Procedia Computer Science, 161: 677-686.
2. Ayub, S. and Jafri, Y.Z. (2020). Comparative study of an ANN-ARIMA hybrid model for predicting Karachi stock price. American Journal of Mathematics and Statistics, 10(1): 1-9.
3. Box, G. E. P.; Pierce, D. A. (1970). "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models". Journal of the AmericanStatisticalAssociation. 65 (332):1509-
1526. doi:10.1080/01621459.1970.10481180. JSTOR 2284333.
4. Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994), Time Series Analysis: Forecasting and Control (3rd ed.). Holden-Day, San Francisco.
5. Cao, L.J. and E.H. Tay, 2001. Support vector with adaptive parameters in financial time series forecasting. IEEE Trans. Neural Network, 14:1506-1518
6. Chadsuthi,S., Modchang, C., Lenbury, Y.,Iamsirithaworn, S. and Triampo, W.(2012). Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time - series and ARIMAX analyses, Asian Pacific Journal of Tropical Medicine,539 -546.
7. Flake, G.W. and S. Lawrence, 2002. Efficient SVM regression training with SMO. Machine learning, 46:271-290
8. G. M. Ljung; G. E. P. Box (1978). "On a Measure of a Lack of Fit in Time Series Models". Biometrika. 65 (2): 297–303. doi:10.1093/biomet/65.2.297
9. Jha, G.K. and Sinha, K. (2014). Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in India. Neural Comput. Appl., 24(3).
10. Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). Forecasting: Methods and Applications (3rd ed.). Wiley, Chichester.
11. Padhan, P.C. (2012), Application of ARIMA model for forecasting agricultural productivity in India. J. Agric. Soc. Sci., 8:50-56

12. Rathod, Santosha., Mishra, G.C. and Singh, K. N. (2017). Hybrid time series models for forecasting Banana production in Karnataka State, India. Journal of the Indian Society of Agricultural Statistics., 71(3): 193-200.
13. Ray, S., Bhattacharyya, B. (2020). Statistical modeling and forecasting of arima and arimax models for food grains production and net availability of India. Journal of Experimental Biology and Agricultural Sciences, 8(3): 296 – 309
14. Vapnik, V.N. (1995). The nature of Statistical learning Theory. 1st edn., Springer-Verlog, New York, ISBN: 0-387-94559-8
15. Zhang, G. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.
16. Zhao, C. Y., H.X. Zhang, M.C. Liu, Z. D. Hu and B.T. Fan. (2006). Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology, 217: 105-119