Indian Journal of Agricultural Sciences **88** (12): 1909–13, December 2018/Article

Improved ARIMAX modal based on ANN and SVM approaches for forecasting rice yield using weather variables

WASI ALAM[1], MRINMOY RAY[2], RAJEEV RANJAN KUMAR[3], KANCHAN SINHA[4], SANTOSHA RATHOD[5] and K N SINGH[6]

ICAR-Indian Agricultural Statistics Research Institute, Pusa, New Delhi 110 012

Received: 2 April 2018; Accepted: 9 July 2018

ABSTRACT

An effort has been made to get precise forecast of rice yield through ARIMAX and proposed hybrid models using weather variables. In this article, two hybrid approaches like ARIMAX-ANN and ARIMAX-SVM have been proposed. Firstly, ARIMAX model was fitted for the considered time series data. Rice yield along with weather variables of Aligarh district of Uttar Pradesh have been considered to evaluate the forecasting performance of the proposed hybrid models. The residuals obtained from the fitted model which exhibit nonlinear pattern were fitted employing ANN and SVM. Using the fitted yield values through the hybrid approaches via ANN and SVM, MAPE under ARIMAX (0,1,1)-ANN and ARIMAX (0,1,1)-SVM are estimated to be 0.37 and 1.11, respectively, as compared to 12.18 under ARIMAX (0,1,1) model. Based on the results obtained, we infer that although performance of proposed ARIMAX-SVM and ARIMAX-ANN models are close to each other but much superior to the conventional ARIMAX model for the considered data set. Performance of hybrid ARIMAX model is found to be quite encouraging. Yield has also been forecasted up to 2020 on the basis of forecasted rainfall using ARIMAX (0,1,1) model.

Key words: ANN, ARIMAX, Hybrid ARIMAX, SVM

Among foodgrains, rice is the most important crop of the developing world and the staple food for more than 60% of the Indian population. Projection of rice demand by 2030 mentioned in vision 2030 of National Rice Research Institute (Adhya *et al*. 2011) has been computed on the basis of fixed historical growth rate. This approach is quite adhoc and having no sound statistical foundation. In order to get more reliable future crop yield/production forecast, we need more precise time series forecast. Traditionally, classical ARIMA model (Box *et al.* 2007) has been widely used for short term time series forecasting, but not for long term forecasts due to the convergence of the autoregressive part of the model to the mean of the time series. Moreover, this approach does not explain the nonlinear component of residuals obtained through ARIMA model. Rice is a rainfed crop and due to climate change rice yield may be tremendously influenced by weather variables. Hence, it is quite interesting to assess the impact of time series weather variables on yield and consequently rice production. The ARIMAX model is a generalization of ARIMA model which is capable of incorporating an external input variable. Hyndman (2010) preferred to call ARIMAX as regression with ARIMA errors. An effort has been made to find out suitable ARIMAX model using weather variables for rice yield data. ANN and SVM (Kim 2003, Sapankevych and Sankar 2009) techniques have been applied on the residuals obtained through the selected ARIMAX model. Fitted residuals obtained through ANN and SVM have been used to correct the fitted yield of ARIMAX model. We have tried to improve the performance of ARIMAX on the similar line of Zhang (2003). Here, we have only improved the performance of ARIMAX model through ANN and SVM techniques rather than long term forecast. Work on long term forecast using hybrid ARIMAX model will be the future work. Selected ARIMAX model has been used for forecasting yield up to 2020 using weather variables as exogenous variables. Paul and Sinha (2016) compared ARIMAX with NARX for forecasting crop yield. For other works of authors, one may refer to Paul *et al.* (2011, 2014) and Naveen *et al.* (2012) and Ray *et al.* (2016).

MATERIALS AND METHODS

Data on rice yield of Aligarh district of Uttar Pradesh has been taken from Directorate of Economics and Statistics from the year 1975 to 2013. Weather data on mean maximum temperature in degrees C, mean minimum temperature in degrees C and total weekly rainfall in mm in 24 hr ending

[1]Senior Scientist (e mail: wasi@iasri.res.in, mw.Alam@icar.gov.in), [2]Scientist (e mail: mrinmoy.ray@icar.gov.in), [3]Scientist, [4]Scientist, [5]Scientist, [6]Principal Scientist (e mail: knsingh@iasri.res.in)

0830 hr IST have also been taken from India, Meteorological Department, Pune for the corresponding period for Aligarh district. We have considered rice crop which is *kharif* crop in Uttar Pradesh and rice yield is completely dependent on monsoon, hence, we are focused mainly on weather information from the period of June to September of each year. ARIMAX model is linear in nature and hence does not explain the nonlinearity component. Here, we have tried to improve the performance of ARIMAX model by explaining residuals through machine learning approaches like ANN and SVM on the line of Zhang (2003). In the first phase, the time series is analyzed by using ARIMAX models. In the next phase, the residuals obtained in the previous phase are examined by ANN and then forecast values obtained from the ARIMAX model are summed. The residuals of the linear model will then contain only the nonlinear relationship. Therefore, in the second phase, the ANN and SVM are used to model the nonlinear patterns of ARIMAX residuals. Artificial Neural Networks are flexible computing frameworks for modeling a broad range of nonlinear problems. One significant advantage of the ANN models over other classes of nonlinear models is that ANNs are universal approximators that can approximate a large class of functions with a high degree of accuracy. A typical time delay neural network structure with one hidden layer is denoted by I:Hs:O*l*, where I is the number of nodes in input layer, s denotes the logistic sigmoid transfer function, O denotes number of nodes in the output layer and *l* indicates linear transfer function.

The ARMAX model is a generalization of ARMA model which is capable of incorporating an external input variable. ARIMA model is extended into ARIMA model with exogenous variable X, called ARIMAX (p, d, q). Let the time series be denoted by y_1, y_2, ..., y_n and we assume the series to be stationary, hence, we consider ARMA model. First, we define an ARMA (p, q) model with no covariates:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where ε_t is a white noise process (i.e. identically independently distributed with mean=0).

An ARMAX model simply adds in the covariate on the right hand side:

$$y_t = \beta x_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where β is a covariate at time t and β is its coefficient.

If we write the model using backshift operators, the ARMAX model is given by

$$\phi(B) y_t = \beta x_t + \theta(B) \varepsilon_t$$

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

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

We note that autoregressive coefficients get mixed up with both the covariates and the error term. Hyndman (2010) preferred it to call regression with ARMA errors. Regression models with ARMA errors is defined as

$$y_t = \beta x_1 + \eta_t$$

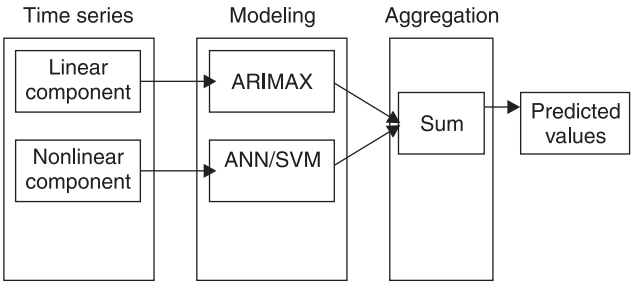
$$\eta_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

In this case, the regression coefficient has its usual interpretation. Using backshift operators, this model can be written as

$$y_t = \beta x_t + \frac{\theta(B)}{\phi(B)} \varepsilon_t$$

when the data is non-stationary, for the ARIMA errors, we simply replace $\phi(B)$ with $\Delta^d \phi(B)$ where $\Delta = 1\text{-}B$ denotes the differencing operator. Differencing of y_t and x_t is required before fitting the model with ARMA errors. Hence, differencing all variables is necessary because estimation of a model with non-stationary errors is not consistent and can lead to spurious regression. The first step in building an ARMAX model consists of identifying a suitable ARMA model for the endogenous variable. The ARMAX model concept requires to test for stationarity of exogenous variable before modeling. Nonlinear least square estimation procedure is employed to estimate the parameters of ARMAX model. In the above model, crop yield has been considered as dependent variable (Y) while minimum temperature, maximum temperature and rainfall as exogenous variables (X). To this end, forecast of covariates using hybrid time domain approaches with ANN and SVM have been utilized in the fitted ARIMAX. Performance of developed forecast models have been thoroughly examined.

The proposed forecasting approaches can be graphically represented as below:



The main strategy of this approach is to model the linear component L_t and nonlinear component N_t separately by different models. Initially, an ARIMAX model is employed to fit the linear component. Let the prediction series provided by ARIMAX model denoted as $\hat{L}_t$. In the second step, the residuals denoted as e_t which are nonlinear in nature are predicted. The residuals can be estimated by subtracting the predicted value y_t from actual value of the considered time series y_t.

$$e_t = y_t - \hat{L}_t$$

Now the residuals are predicted employing an ANN/SVM model (Vapnik 2000). Let the prediction series provided by ANN model denoted as $\hat{N}_t$. Eventually, the predicted linear and nonlinear components are combined to generate aggregate prediction.

$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

Above will be the fitted yield through hybrid approach and using these fitted values, we compute MAPE for

comparing the forecast accuracy of hybrid ARIMAX approach with the traditional ARIMAX model.

RESULTS AND DISCUSSION

Sequence chart for the yield has been shown in Fig 1. From Fig 1, it is obvious that data is suitable for Box- Jenkins linear time series techniques. Since, yield of rice is dependent on monsoon, hence, we are focused mainly on weather information from June to September. ARIMAX model has been applied for modelling and forecasting of rice yield using exogenous variables like rainfall, minimum temperature and maximum temperature for the Aligarh district of Uttar Pradesh for the years 1975 to 2010 as a training set of the data. On the basis of minimum values of goodness of fit values (Table 1), ARIMAX (0,1,1) model with rainfall was found to be the suitable one as AIC and BIC are 7.02 and 11.688, respectively. P-values and parameters estimate of ARIMAX (0,1,1) with rainfall as exogenous variable are mentioned in Table 2. P-values for MA (1,1) term and rainfall are found to be <0.0001 and 0.0312, respectively, which are significant at 5% level. MAPE under ARIMAX (0,1,1) model with rainfall as exogenous variable is estimated to be 12.18% as compared to 17.68% under usual ARIMA (2,1,0) model without covariates. From the Fig 2, residuals are white noise. Hence, we found the ARIMAX (0,1,1) model as suitable model under rainfall as exogenous variable. ANN approach was applied on the residuals of ARIMAX (0,1,1) for modeling and forecasting of the residuals. ANN model with 06:04s:1l has been identified as suitable model as this model has minimum values of MAE, i.e 0.005 and 0.18 under training and testing, for more details of the results, please see Table 3. Using 06:04s:1l model, estimated the fitted values of residuals and these fitted residuals were used to correct the fitted values of yield obtained through ARIMAX(0,1,1) model and eventually got the fitted values under hybrid ARIMAX (0,1,1)-ANN. MAPE under this hybrid approach is estimated to be 0.37% as compare to 12.18% under simply ARIMAX (0,1,1) model for training set of the data. SVM was also applied on the residuals of the ARIMAX (0,1,1) model and got fitted residuals by SVM and the fitted values of yield obtained by ARIMAX (0,1,1) model were corrected by the fitted residuals of SVM. The MAPE for the corrected yield using fitted residuals of SVM is estimated to be 1.11. For validation, Forecast values through ARIMAX (0,1,1), ARIMAX (0,1,1)- SVM and ARIMAX(0,1,1)-ANN models are given in Table 4 along with the actual values. The significant reduction in MAPE from 12.18 (for ARIMAX (0,1,1)) to 0.37 and 1.11 through ARIMAX (0,1,1)-ANN and ARIMAX (0,1,1)-SVM, respectively, indicates a significant improvement in the performance of ARIMAX model using machine learning techniques. Hence, hybrid ARIMAX model using ANN and SVM approaches may be recommended for more

Table 1 Goodness of fit

Model	AIC	SBC
ARIMAX $(0,1,1)_{Rainfall}$	7.021955	11.688
ARIMAX $(2,1,0)_{Rainfall}$	17.71695	23.93835
ARIMAX $(0,1,1)_{Max.Temp}$	10.22489	14.89093
ARIMAX $(2,1,0)_{Max.Temp}$	19.01006	25.23145
ARIMAX $(2,1,0)_{Min.Temp}$	18.77149	24.99289
ARIMAX $(1,1,0)_{Rainfall, Min.Temp, Max. Temp}$	21.63915	29.41589
ARIMAX $(2,1,0)_{Rainfall, Min.Temp}$	19.36	27.13516

Table 2 Parameter estimates of the model

Parameter	Estimate	Standard error	t value	Approx Pr > |t|	Lag	variable
MU	-0.08809	0.05781	-1.52	0.1374	0	Yield
MA1,1	0.99900	0.13219	7.57	<.0001	1	Yield
NUM1	0.0002306	0.0001023	2.25	0.0312	0	Rainfall

Table 3 Forecast accuracy

Model	No. of parameters	MAE for training	MAE for testing
1:2s:1*l*	7	0.206	0.12
1:4s:1*l*	13	0.166	0.12
1:6s:1*l*	19	0.166	0.13
1:8s:1*l*	25	0.151	0.16
1:10s:1*l*	31	0.141	0.27
2:2s:1*l*	9	0.152	0.15
2:4s:1*l*	17	0.104	0.41
2:6s:1*l*	25	0.07	0.09
2:8s:1*l*	33	0.038	0.3
3:2s:1*l*	11	0.136	0.15
3:4s:1*l*	21	0.057	.24
3:6s:1*l*	31	0.023	0.22
4:2s:1*l*	13	0.099	0.29
4:4s:1*l*	25	0.025	0.29
5:2s:1*l*	15	0.096	0.2
5:4s:1*l*	29	0.014	0.2
6:2s:1*l*	17	0.076	0.19
6:4s:1*l*	33	0.005	0.018
7:2s:1*l*	19	0.071	0.16
8:2s:1*l*	21	0.028	0.18
9:2s:1*l*	23	0.025	0.36

Table 4 Forecast of rice yield (tonnes/ha) for validation of the model

Year	Actual Y	ARIMAX (0,1,1)	ARIMAX (0,1,1) -ANN	ARIMAX (0,1,1) -SVM
2006	2.06	2.0138	1.997	2.0338
2007	2.01	2.078	1.991	2.008
2008	2.15	2.1052	2.156	2.1252
2009	1.73	2.1163	2.079	1.7463
2010	2.04	2.132	2.185	2.042

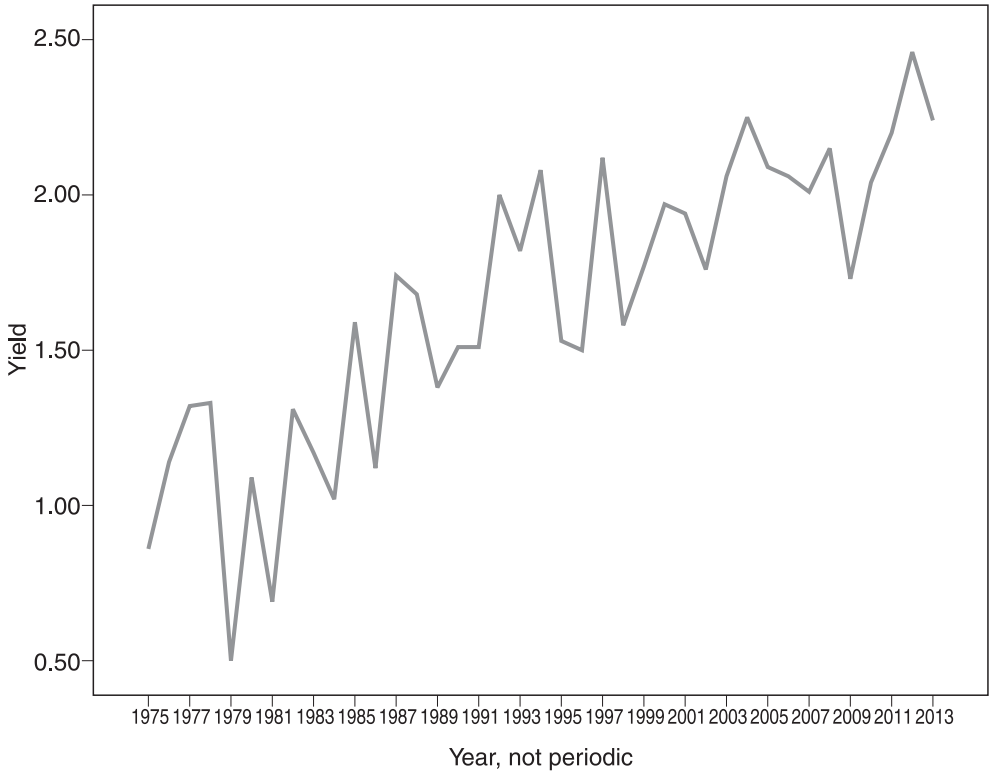


Fig 1 Sequence charts for yield (tonnes/ha).

precise forecast of yield up to the desired forecast horizon and eventually the forecasted yield values may be utilized for computing crop production for estimating per capita availability for the population in long term.

For out of sample forecast of yield through the selected ARIMAX model, we forecasted the rainfall using ANN approach. In order to get forecast values of rainfall (important variable identified on the basis of goodness of fit measures), we applied artificial neural network model on rainfall data and found 05:04s:1*l* as appropriate model with the MAE for training and testing 1.73 and 21.79, respectively and are given in Table 5. Using 05:04s:1*l* model, we got out of sample forecast of rainfall by 2020. Using the forecasted rainfall values over the selected ARIMAX model, rice yield has been forecasted upto 2020 along with 95 % confidence intervals and are presented in Table 6. R and SAS software packages have been used for the purpose.

Table 5 ANN models for rainfall data

Model	No. of parameters	MAE for training	MAE for testing	MAPE training	MAPE testing
1:2s:1l	7	131.46	119.84	23.88	26.81
1:4s:1l	13	112.77	160.05	21.1	32.02
1:6s:1l	19	94.62	246.73	17.84	32.321
1:8s:1l	25	83.84	257.79	15.81	33.25
1:10s:1l	31	108.82	319	14.72	37.39
2:2s:1l	9	108.9	101.32	20.05	22.99
2:4s:1l	17	59.72	124.23	22.99	29.49
2:6s:1l	25	28.05	819.99	5.33	507.33
2:8s:1l	33	15.71	436.28	2.92	96.6
3:2s:1l	11	82.81	146.54	16.81	23.02
3:4s:1l	21	30.93	169.85	6.47	27.53
3:6s:1l	31	8.89	329.29	1.77	47.36
4:2s:1l	13	70.47	244.85	14.17	40.69
4:4s:1l	25	20.1	150.92	4.41	20.49
5:2s:1l	15	63.78	185.91	12.67	29.02
5:4s:1l	29	8.04	133.69	1.73	21.79
6:2s:1l	17	48.83	293.15	9.42	41.21

Conclusion

On the basis of above findings, we may conclude that if a time series variable is caused by some exogenous variables like weather variables, in that case, ARIMAX model must be preferred over simply ARIMA model. For further improvement of the performance of ARIMAX model, application of hybrid ARIMAX using ANN and SVM is recommended on the basis of significant reduction in MAPEs. The performance of the hybrid models can further be improved by applying some other techniques. In future work, apart from weather variables as exogenous variables other variables like fertilizer dose, irrigation level etc. may be considered for better performance and this work may be extended on other crops too including horticultural crops.

Table 6 Out of sample forecast of rice yield (in tonnes/ha) using ARIMAX (0,1,1) model based on forecast values of rainfall

Year	Forecast	Actual	Lo 95	Hi 95
2011	2.090	2.2	1.552	2.627
2012	2.142	2.46	1.568	2.716
2013	1.936	2.24	1.328	2.544
2014	2.008		1.368	2.649
2015	2.043		1.372	2.714
2016	2.007		1.307	2.708
2017	1.985		1.256	2.713
2018	1.972		1.216	2.728
2019	1.902		1.120	2.684
2020	1.913		1.106	2.720

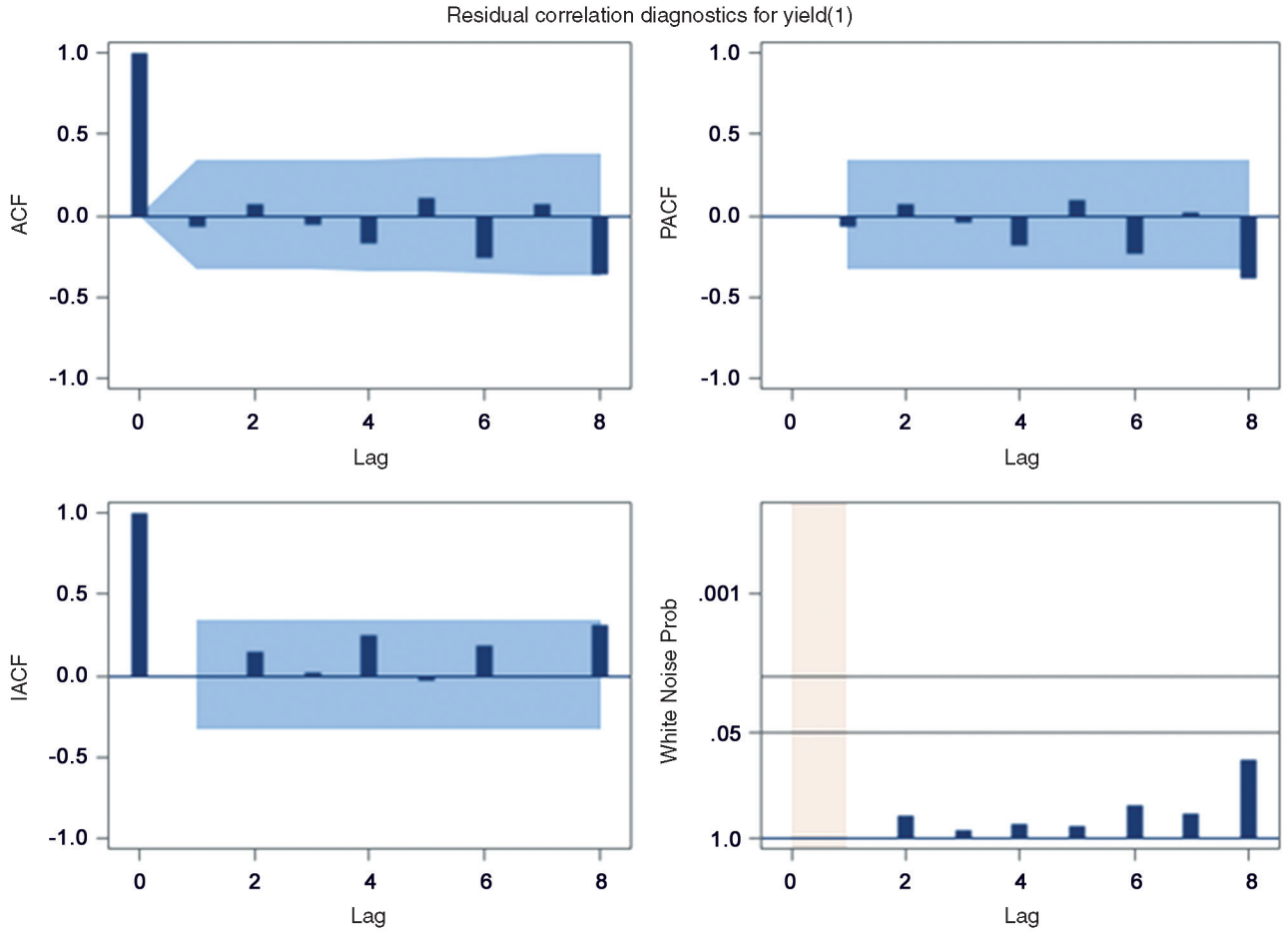


Fig 2 White noise test of residuals.

ACKNOWLEDGEMENT

Authors would like to thank ICAR-Indian Agricultural Statistics Research Institute for funding the project entitled Development of Hybrid Time Series Models using Machine Learning Techniques for Forecasting Crop Yield with Covariates, from which this paper is drawn.

REFERENCES

Adhya T K, Kar P and Sinha S K. 2011. Vision 2030. ICAR-Central Rice Research Institute. [http://www.crri.nic.in/crri_vision2030_2011.pdf]

Box G E P, Jenkins G M and Reinsel G C. 2007. *Time-Series Analysis: Forecasting and Control*, 3rd edition. Pearson education, India.

Hyndman R J. 2010. Forecasting, R, statistics. https://robjhyndman.com/hyndsight/arimax/

Kim K J. 2003. Financial time series forecasting using support vector machines. *Neurocomputing* **55:** 307–19.

Naveen N C, Kumar D, AlamW, Chaubey R , Subramanian S and Raman R. 2012. A model study integrating time dependent mortality in evaluating insecticides against *Bemisia tabaci* (Hemiptera: Aleyrodidae). *Indian Journal of Entomology* **74**(4): 384–8.

Paul R K and Sinha K. 2016. Forecasting crop yield: A comparative assessment of ARIMAX and NARX model. *RASHI* **1** (1): 77–85.

Paul R K, Alsam W and Paul A K. 2014. Prospects of livestock and dairy production in India under time series framework. *Indian Journal of Animal Sciences* **84(4)**: 462–6.

Paul A K, Alam W and Singh P. 2011. Average linkage method for clustering rice producing states of india. *Indian Journal of Agricultural Sciences* **81**(8): 75–8.

Ray M, Rai A, Ramasubramanian V and Singh K N. 2016. ARIMA-WNN hybrid model for forecasting wheat yield time-Series data. *Journal of the Indian Society of Agricultural Statistics* **70(1)**: 63–70.

Sapankevych N I and Sankar R. 2009. Time series prediction using using support vector machines: A survey. *IEEE Computational Intelligence Magazine* **4(2):** 24–38.

Vapnik V. 2000. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Zhang G P. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neuro-computing* **50:** 159–75.