Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 76(2) 2022 67–77

Autoregressive Integrated Moving Average models for Sugarcane Yield Estimation in Haryana

Pushpa, Aditi, Chetna and Urmil Verma

CCS Haryana Agricultural University, Hisar

Received 16 June 2022; Revised 21 September 2022; Accepted 26 September 2022

SUMMARY

Crop yield models are abstract presentation of the interaction of crop with its environment and can range from simple correlation of yield with a finite number of variables to the complex statistical models with predictive end. Autoregressive Integrated Moving Average **(**ARIMA) models have been fitted for the yield data of sugarcane crop in Karnal, Ambala and Kurukshetra districts of Haryana. The crop yield data of the past four/five decades have been used for the model building and the forecast values are obtained for the years 2016-17 to 2020-21. After experimenting with different lags of the moving average and autoregressive processes; ARIMA (0,1,1) for Karnal, Ambala and Kurukshetra districts have been fitted for sugarcane yield forecasting. The overall results indicates that the percent relative deviations of the forecast yield(s) from the real-time yield(s) are within acceptable limits and favors the use of ARIMA models to get short-term forecast estimates.

Keywords: Autocorrelation, Partial autocorrelation, Differencing, Stationarity and invertibility.

1. INTRODUCTION

Forecasting of crop yield is a formidable challenge. Crop yield models are abstract presentation of the interaction of the crop with its environment and can range from simple correlation of yield with a finite number of variables to the complex statistical models with predictive end. Various organizations in India and abroad are engaged in developing methodology for pre-harvest forecasting of crop yield using various approaches. In the standard regression analysis, the various observations within a single series are assumed to be statistically independent. However, with most time-series data, this assumption may not hold true. Therefore, the standard regression analysis is generally not adequate for forecasting time series data as the observations in the series may not be statistically independent. The Box-Jenkins (1976) methodology is a powerful tool for time-series analysis, when the time-sequenced observations in a data series may be statistically dependent or related to each other. The study of yield trends for principal crops in India has been made by several research workers. Panse (1959, 1964) in a series of papers studied the trends in yields of rice and wheat with a view to compare the yield rates during the plan period with that of the pre-plan period. Padhan (2012), Debnath *et al.* (2013), Prabakaran *et al.* (2013), Paul and Ghosh (2013), and Ali *et al.* (2015) have used time series analysis for crop yield forecasting. Paul *et al.* (2013) employed SARIMA model for modelling and forecasting of monthly export of meat and meat products from in India. Hossain and Abdulla (2015) have obtained the forecast of sugarcane production in Bangladesh using Box-Jenkins ARIMA modelling. Paul (2015) used the autoregressive integrated moving average with exogenous variable-Generalized autoregressive conditional heteroscedastic (ARIMAX-GARCH) model to describe volatility data by adding exogenous variables in the mean-model. Paul *et al.* (2015) have applied autoregressive fractionally integrated moving average (ARFIMA) model for modelling and forecasting of daily retail price of pigeonpea (Cajanascajan) in Karnal and GARCH model for price volatility in food commodities in India. Vishwajith *et al.* (2016) have applied univariate ARIMA

Corresponding author: Pushpa
E-mail address: pushpa841991@gmail.com

models for sugarcane area, production and productivity estimation. Fattah *et al.* (2018) have used ARIMA model for forecasting of demand. Pardhi *et al.* (2018) have used ARIMA model for forecasting of monthly price of mango. Wadhawan and Singh (2019) examined the different volatility estimators and determined the efficient volatility estimator. In this study, the emphasis has been given to forecast future values on the basis of previous time-series observations. In accordance with the objective formulated, 'ARIMA models for sugarcane yield forecasting in Haryana' has been fitted to see the forecasting performance of the developed ARIMA models.

2. DATA DESCRIPTION AND MODELING PROCEDURE OF ARIMA (P,D,Q)

The sugarcane yield for the period 1966-67 to 2020-21 of Karnal and Ambala districts and 1972-73 to 2020-21 of Kurukshetra district were compiled from the Statistical Abstracts of Haryana. The ARIMA models were developed using the sugarcane yield data for the period 1966-67 to 2015-16 of Karnal and Ambala districts and 1972-73 to 2015-16 of Kurukshetra district and sugarcane yield forecasts on the basis of fitted models were done for the years 2016-17 to 2020-21.

The general functional form of ARIMA (p,d,q) model (used) is :

$\varphi_p(B)\Delta^d Y_t = C + \theta_q(B)a_t$

where,

Y_t = Variable under forecasting

B = Lag operator

a = Error term ($Y_t - \hat{Y}_t$, where $\hat{Y}t$ is the estimated value of Y_t)

t = the time subscript

$\varphi_p(B)$ = Non-seasonal AR

$(1-B)^d$ = Non-seasonal difference

$\theta_q(B)$ = Non-seasonal MA

3. ACCURACY STATISTICS

The forecasting performance of the ARIMA model is examined in terms of accuracy measures such as Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Relative deviation percentage (RD%) are given below.

$$RMSE = \sqrt{\frac{1}{n}\left(observed - predicted\right)^2}$$

$$MPAE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{observed - predicted}{observed}\right|$$

$$RD\% = \frac{observed - predicted}{observed} \times 100$$

4. RESULTS AND DISCUSSION

4.1 Identification

Identification involves the determination of appropriate orders of AR and MA polynomials i.e. the values of p and q. The orders were determined from the autocorrelation functions (acfs) and partial autocorrelation functions (pacfs) of the stationary series. The plotting of acfs for all the districts under consideration shown in Fig. 1, indicated that the acfs declining gradually imply non-stationarity. Differencing of order one was enough for getting an appropriate stationary series for all the districts. The pacfs showed the presence of one significant spike at lag one, indicating that the series may have one order of AR component are shown in Fig. 2. The non-stationary data series of all the districts were transformed into stationary series by the first differencing of the original data series. However, for Ambala district, the log transformation was also tried to meet the stationarity

Table 1. Selection criteria values of ARIMA models considered for all the districts

District	Model	Model fit statistic(s)			
		RMSE	MAPE	MAE	BIC
Karnal	ARIMA (1,1,1)	6.19	8.21	4.95	3.87
	ARIMA (0,1,1)	6.14	8.19	4.59	3.77
	ARIMA (1,1,0)	7.02	9.11	5.19	4.04
Ambala	ARIMA (1,1,1)	5.73	8.48	4.32	3.71
	ARIMA (0,1,1)	5.69	8.47	4.28	3.63
	ARIMA (1,1,0)	6.14	9.97	5.04	3.77
Kurukshetra	ARIMA (1,1,1)	5.73	8.49	4.32	3.71
	ARIMA (0,1,1)	5.70	8.43	4.12	3.70
	ARIMA (1,1,0)	7.12	9.88	5.36	4.08

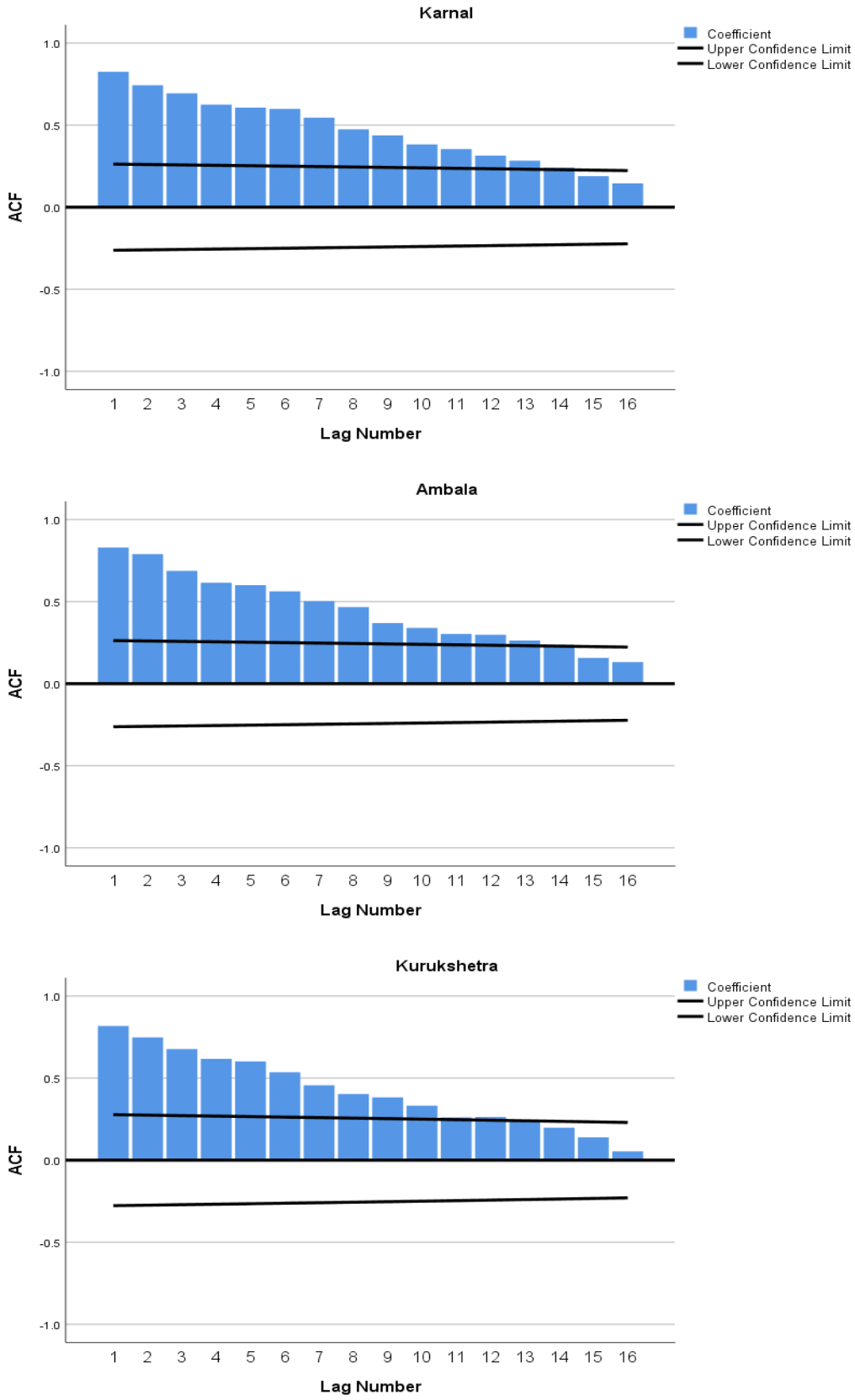

Fig. 1. Autocorrelations of sugarcane yield for all the districts

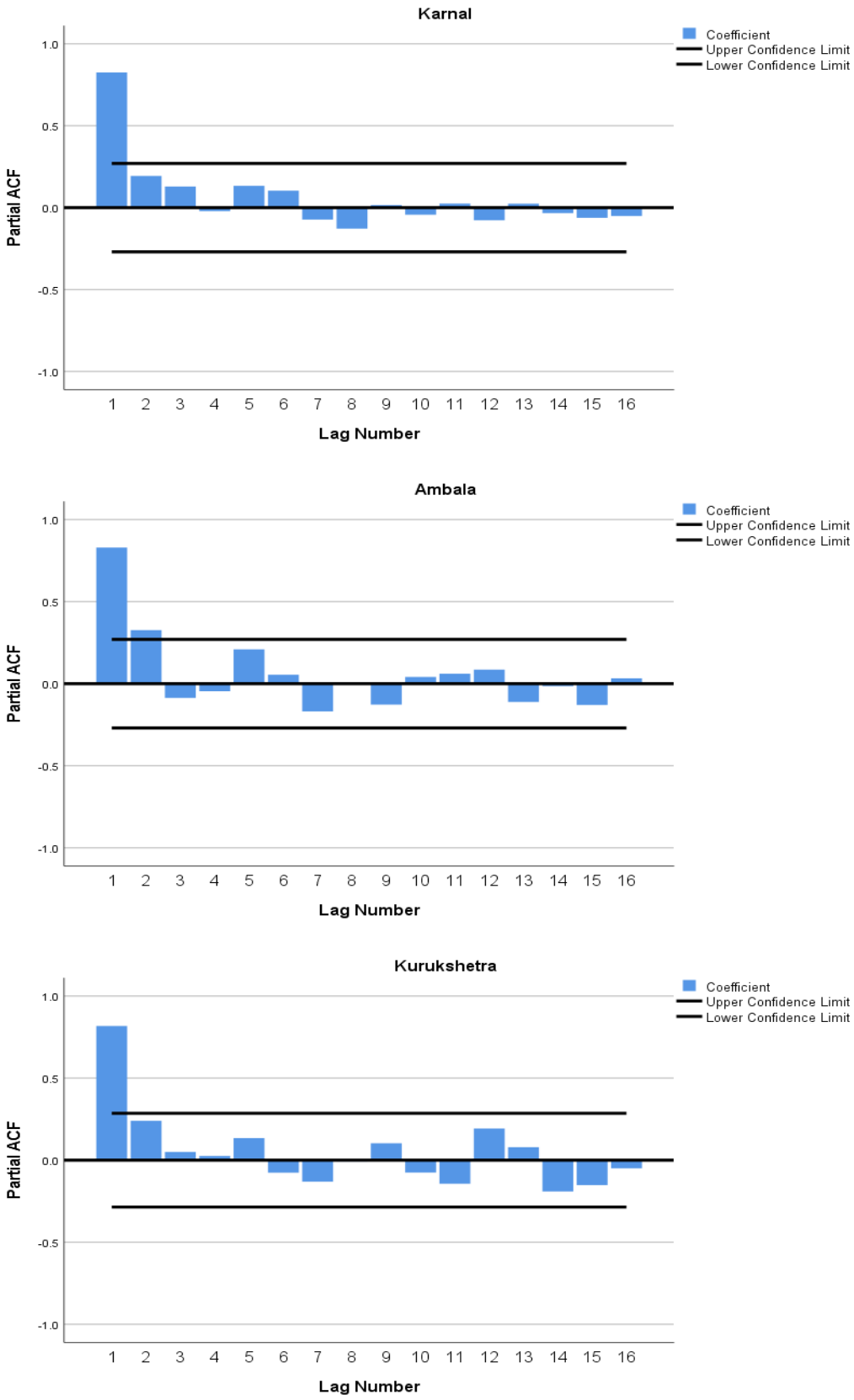

Fig. 2. Partial autocorrelation of sugarcane yield for all the districts

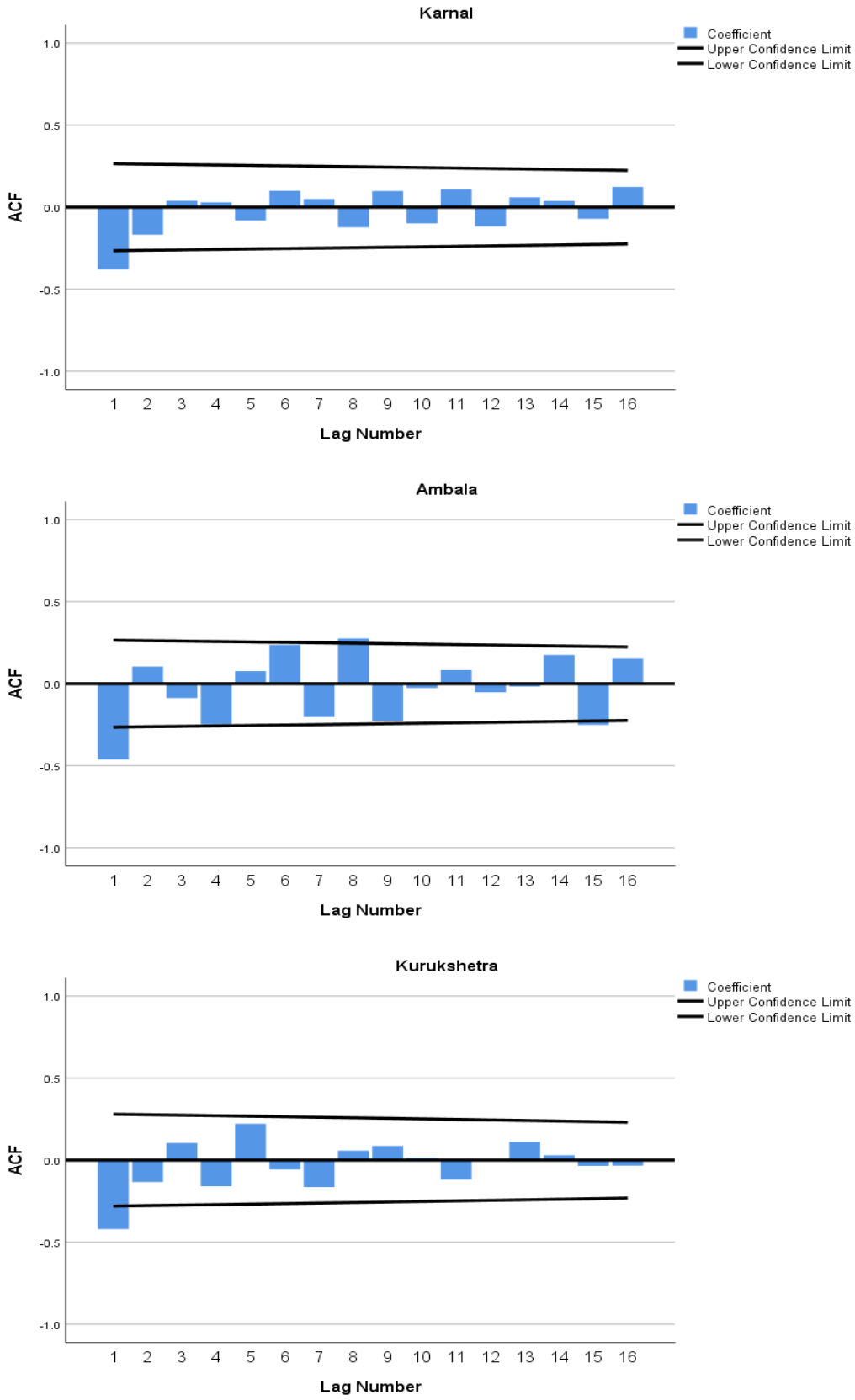

Fig. 3. Autocorrelation of sugarcane yield after 1st differencing for all the districts

requirement but couldn't give any clear view in this regard. Fig. 3 indicate that differencing of order one i.e. d=1 was enough for getting an approximate stationary series in all the districts. Tentative ARIMA (1,1,1), ARIMA (0,1,1) and ARIMA (1,1,0) were fitted to estimate the district-level sugarcane yield for Karnal, Ambala and Kurukshetra districts in identification stage.

4.2 PARAMETER ESTIMATION

The models ARIMA (1,1,1), ARIMA (0,1,1) and ARIMA (1,1,0) were considered in the identification stage and parameter estimation was carried out using a non-linear least squares (NLS) approach. Least squares estimates give the smallest sum of squared residuals. Linear least squares may be used to estimate only pure AR models. All other models require a non-linear least squares method. Model are preferred with a smallest RMSE, MAPE, MAE and BIC, since it tends to produce forecast with smaller error variance. Among the several NLS methods, the one most commonly used to estimate ARIMA models is known as Marquardt's compromise. Marquardt has designed a powerful algorithm where some preliminary estimates are chosen first and then the computer program refines them iteratively so as to minimize the sum of squared residuals. Marquardt algorithm (1963) was used to minimize the sum of squared residuals. Log Likelihood, Akaike's Information Criterion, AIC (1969), Schwarz's Bayesian Criterion, SBC (1978) and residual variance decided the criteria to estimate AR and MA coefficients in the model(s). Parameter estimates for the fitted models shown in Table 2, ARIMA (0,1,1) and ARIMA (1,1,0) model parameter is significant. Absolute value of parameter estimation is less than one (needed for convergence) and also satisfies the stationarity and invertibility conditions under ARIMA models shown in Table 3.

It is clear from the Table 3 that the invertibility condition is satisfied because absolute value of MA coefficient in all the districts are less than one.

4.3 Diagnostic Checking

The model verification was concerned with checking the residuals to see if they contained any systematic pattern which can be removed to improve the chosen ARIMA models. Approximate t-values were calculated for residual autocorrelation coefficients using Bartlett's approximation for the standard error of the estimated autocorrelations. All chi-Squared statistic(s) in this concern were calculated using the Ljung-Box (1978) formula as has been shown in Table 4. The graphical Fig. 4 and 5 showed that none of the residual acfs in any of the districts were significantly different from zero at a reasonable level and residuals like approximately normal. This ruled out any systematic pattern in the residuals.

After experimenting with different lags of the moving average and the autoregressive processes,

Table 2. Parameter estimates of the fitted models for sugarcane yield of all the districts

Districts	Model		Estimate	Std. error	t-value	P-value
Karnal	ARIMA (1,1,1)	AR	0.07	0.18	0.43	0.66
		MA	0.82	0.12	7.06	<0.01
	ARIMA (0,1,1)	MA	0.79	0.09	8.67	<0.01
	ARIMA (1,1,0)	AR	-0.39	0.13	-3.02	0.04
Ambala	ARIMA (1,1,1)	AR	0.10	0.17	0.61	0.54
		MA	0.86	0.09	8.79	<0.01
	ARIMA (0,1,1)	MA	0.83	0.09	9.71	<0.01
	ARIMA (1,1,0)	AR	-0.49	0.12	-4.12	<0.01
Kurukshetra	ARIMA (1,1,1)	AR	0.05	0.16	0.33	0.74
		MA	0.99	2.49	0.40	0.69
	ARIMA (0,1,1)	MA	0.94	0.05	16.30	<0.01
	ARIMA (1,1,0)	AR	-0.41	0.13	-3.09	<0.01

Table 3. Stationarity and invertibility requirements for AR and MA coefficients

District	Model		Stationarity	Invertibility
Karnal	ARIMA (0, 1, 1)	MA	*	0.79
	ARIMA (1, 1, 0)	AR	0.39	**
Ambala	ARIMA (0, 1, 1)	MA	*	5.69
	ARIMA (1, 1, 0)	AR	6.14	**
Kurukshetra	ARIMA (0, 1, 1)	MA	*	5.70
	ARIMA (1, 1, 0)	AR	7.12	**

* Stationarity condition is not applicable since the model is MA model
** Invertibility condition is not applicable since the model is AR model

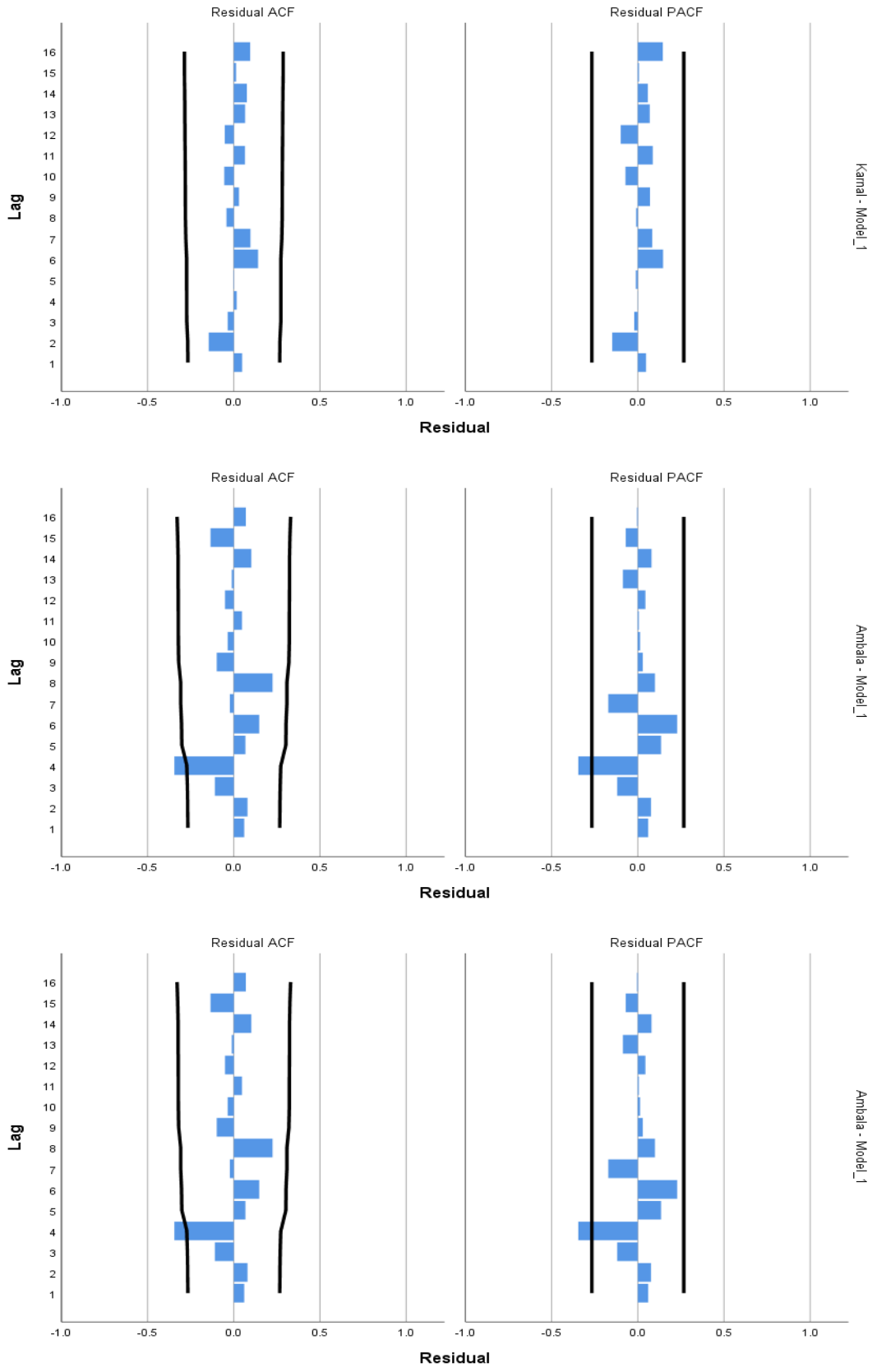

Fig. 4. Residual acfs and pacfs based on fitted ARIMA models for all the districts

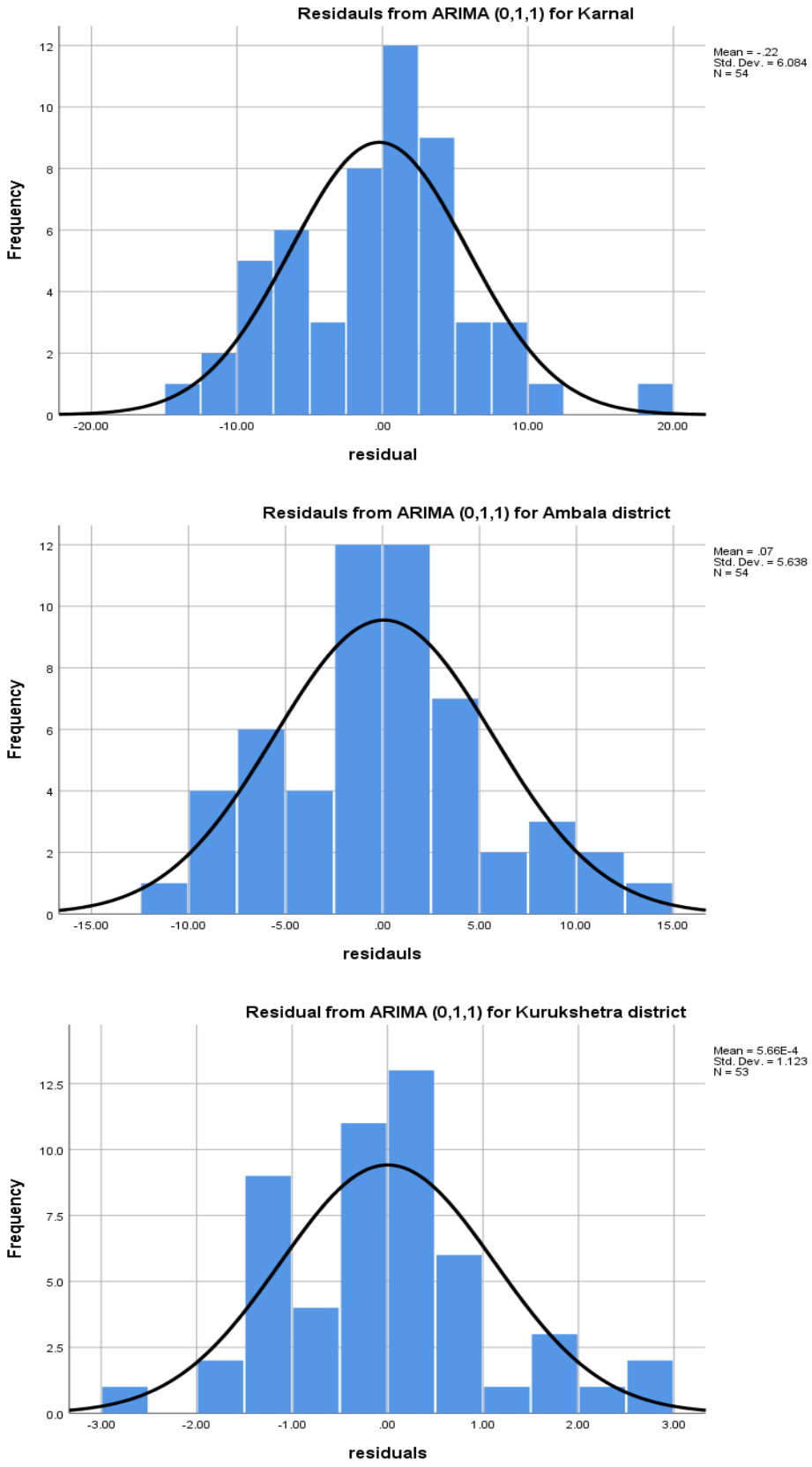

Fig. 5. Residual plots based on fitted ARIMA models for all the districts

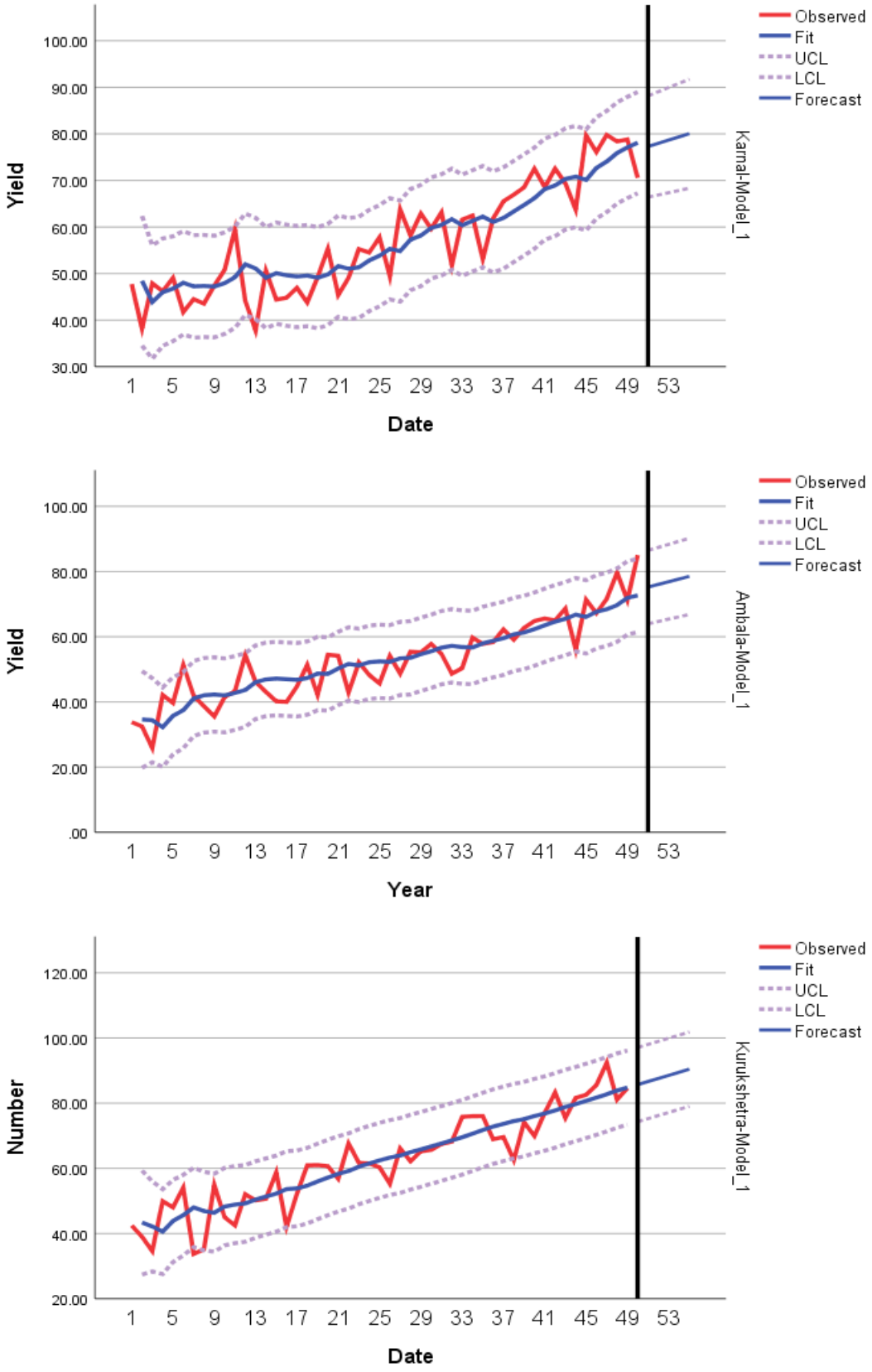

Fig. 6. Actual and predicted sugarcane yield based on ARIMA models for all the districts

ARIMA (0,1,1) and ARIMA (1,1,0) model selected at parameter estimation stagebut at diagnostic checking stage, ARIMA (0,1,1) was found to be the best fit on the basis of non-significant Ljung-box Q statistics i.e. white noise for sugarcane yield estimationfor Karnal, Ambala and Kurukshetra district.

The fitted model ARIMA (0,1,1) for Karnal, Ambala and Kurukshetra districts may be elaborated as below:

$$(1\text{-}B)\, Y_t = (1\text{-}\theta_1 B)a_t$$

$$Y_t\text{-}BY_t = a_t\text{-}\, \theta_1 B a_t$$

$$Y_t = Y_{t\text{-}1} - \theta_1\, a_{t\text{-}1+}\, a_t \qquad (1)$$

The equation 1 is the corresponding forecast equation. The plot of observed and predicted values (Fig. 6) shows that all the fitted models give a good representation of the underlying process.

Table 4. Residual autocorrelations checking based on ARIMA models for all the districts

District	Model	Ljung-box Q statistic		
		Statistic	d.f.	Sig.
Karnal	ARIMA (0,1,1)	6.09	17	0.99
	ARIMA (1,1,0)	34.09	17	0.06
Ambala	ARIMA (0,1,1)	20.60	17	0.24
	ARIMA (1,1,0)	32.36	17	0.04
Kurukshetra	ARIMA (0,1,1)	11.96	17	0.80
	ARIMA (1,1,0)	20.91	17	0.03

The ARIMA (0,1,1) model was used to obtain the sugarcane yield predicted for the periods 2016-17 to 2020-21 and forecast periods from 2021-22 to 2023-24 shown in table 5 and 6 respectively. Finally, a comparison between ARIMA based yield estimates with DOA yield estimates was made in terms of percent relative deviation (RD%), RMSE and MAPE. The results presented in Table 5 indicate that the deviation of the predicted yield from the actual yield, RMSE and MAPE are low for all district, favoring the use of ARIMA models to get short-term forecast estimates.

Table 5. District-level predicted sugarcane yield (q/ha) based on ARIMA models and their percent relative deviations

District/ Model	predicted Year	Observed yield (q/ha)	Estimated yield (q/ha)	Percent relative deviation
Karnal ARIMA (0,1,1)	2016-17	69.6	77.12	-10.80
	2017-18	95	76.45	19.53
	2018-19	93.13	82.43	11.49
	2019-20	83.99	84.88	-1.06
	2020-21	84.23	84.72	-0.58
	RMSE	10.16	MAPE	8.69
Ambala ARIMA (0,1,1)	2016-17	85.54	81.87	4.29
	2017-18	78.13	83.75	-7.19
	2018-19	81.21	87.81	-8.13
	2019-20	72.14	90.17	-24.99
	2020-21	79.79	89.21	-11.81
	RMSE	10.02	MAPE	11.28
Kurukshetra ARIMA (0,1,1)	2016-17	82.56	80.72	2.23
	2017-18	85.57	81.7	4.52
	2018-19	92.43	82.73	10.49
	2019-20	81.09	83.88	-3.44
	2020-21	84.6	84.77	-0.20
	RMSE	4.90	MAPE	4.17

Table 6. Forecast sugarcane yield (q/ha) based on best fitted ARIMA models for all districts

Year	Karnal	Ambala	Kurukshetra
2021-22	85.32	80.70	85.71
2021-23	84.53	81.55	86.65
2023-24	85.28	82.40	87.59

From the analysis of time series data of sugarcane yield, it is inferred that the Box-Jenkins methodology has produced forecast figures which are quite accurate in the sense that the forecast yield(s) compare favorably with the observed yields. In the end, it is emphasized that some of the aspects such as selection of order of differencing, autoregressive and moving average components are highly sensitive to the model results. A proper care must be taken in identifying/generating these figures for the analysis otherwise the results may mislead the decision makers. However, the technique should be used only for taking the short-term forecasts.

REFERENCES

Akaike, H. (1969). Fitting autoregressive models for prediction, *Annals of Institute of Statistical Mathematics*, **21**, 243-247.

Ali, S., Badar, N. and Fatima, H. (2015). Forecasting production and yield of sugarcane and cotton crops of Pakistan for 2013-2030, *Research Article*, **31(1)**, pp 1.

Box, G.E.P. and Jenkins, G.M. (1976). Time series analysis: Forecasting and control, *Holden Day, San Francisco*, 575.

Debnath, M.K., Bera, K. and Mishra, P. (2013). Forecasting area, production and yield of cotton in India using ARIMA model, *Research & Reviews: Journal of Space Science & Technology*, **2(1)**, 16-20.

Hossain, M.M. and Abdulla, F. (2015). Forecasting the sugarcane production in Bangladesh by ARIMA model, *Journal of Statistics Applications & Probability*, **4**(2), 297-303.

Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models, *Biometrika*, **65**, 297-303.

Marquardt, D.W. (1963). An algorithm for least-squares estimation of non-linear parameters, *Society for Industrial and Applied Mathematics*, **2**, 431-41.

Padhan, P.C. (2012). Application of ARIMA Model for Forecasting Agricultural Productivity in India, *Journal of Agriculture & Social Sciences*, **8**, 50–56.

Prabakaran, K., Sivapragasam, C., Jeevapriya, C., and Narmatha, A. (2013). *Golden Research Thoughts*, **3(3)**, 1-7.

Panse, V.G. (1959). Recent trends in the yield of rice and wheat in India, *Indian Journal of Agricultural Economics*, **14**, 11-38

Panse, V.G. (1964). Yield trends of rice and wheat in first two five-year plans in India, *Journal of Indian Society of Agricultural Statistics*, **16(1)**, 1-50.

Schwarz, G. (1978). Estimating the dimension of a model, *The annals of Statistics*, **62**, 461-464

Vishwajith, K.P., Sahu, P.K., Dhekale, B.S. and Mishra, P. (2016). Modelling and forecasting sugarcane and sugar production in India, *Indian Journal of Economics and Development*, **12(1)**, 71-79.

Paul, R.K., Bhardwaj, S.P., Singl, D.R., Kumar, A., Arya, P and Singh, K.N. (2015). Price volatility in food commodities in india - an empirical investigation, *International Journal of Agricultural and Statistical Sciences*, **11**(2), 395-401.

Paul, R.K. and Ghosh, H. (2013). Statistical modelling for forecasting of wheat yield based on weather variables, *Indian Journal of Agricultural Sciences*, **83(2)**, 180-183.

Paul, R.K. (2015). ARIMAX-GARCH-WAVELET model for forecasting volatile data, *Model Assisted Statistics and Applications,* **10(3)**, 243-252.

Paul, R.K., Panwar, S., Sarkar, S., Kumar, G.V.A., Singh, K.N., Farooqi, M.S. and Choudhary, V.K. (2013). Modelling and Forecasting of Meat Exports from India, *Agricultural Economics Research Review*, **26(2)**, 249-255.

Fattah, J. Ezzine, L. Aman, Z., Moussami, H.E. and Lachhab, A. (2018). Forecasting of demand using ARIMA model, *International Journal of Engineering Business Management*, **10**, 1-9.

Pardhi, R., Singh, R. and Pual, R.K. (2018). Price Forecasting of Mango in Varanasi Market of Uttar Pradesh, *Current Agriculture Research Journal*, **6(2)**, 218-224.

Wadhawan and Singh (2019). Estimating and Forecasting Volatility Using Arima Model: A Study on NSE, India, *Indian Journal of Finance, volume,* **13(5)**, 37-51.