Biological Forum – An International Journal **14(4a): 591-596(2022)**

ISSN No. (Print): 0975-1130
ISSN No. (Online): 2249-3239

Hybrid Sarima-Ann Model for Forecasting Monthly Wholesale Price and Arrival Series of Tomato Crop

Pushpa[1*], Joginder Kumar[2] and Vikram[3]
[1]Ph.D. Scholar, Department of Mathematics and Statistics,
CCS HAU, Hisar-125004 (Haryana), India.
[2]Assistant Scientist, Department of Mathematics and Statistics,
CCS HAU, Hisar-125004 (Haryana), India.
[3]Assistant Scientist, Department of Vegetable Science,
CCS HAU, Hisar-125004 (Haryana), India.

(Corresponding author: Pushpa)*
(Received 21 September 2022, Accepted 18 November, 2022)
(Published by Research Trend)

ABSTRACT: Agricultural prices forecasting are the major concern for the policy makers as they directly affect the profitability of farming as an occupation. A hybrid model is considered to be an effective way to improve forecast accuracy. The hybrid model of the linear seasonal autoregressive moving average (SARIMA) and the nonlinear Artificial neural network (ANN) is proposed in this paper for estimating and forecasting the monthly wholesale price and arrival series of tomato crop. The goodness of fit of the model is measured using Akaike information criteria (AIC), root mean square error (RMSE), and mean absolute percentage error (MAPE), while post-sample forecast accuracy is measured using mean absolute error (MAE) and percent standard error of prediction (SEP). The study clearly shows that the hybrid (SARIMA-ANN) model is superior for forecasting the monthly wholesale prices and arrival series of tomato in the Gurugram market. The R (4.1.3) software is used for the analysis.

Keywords: Price and arrival forecasting, MAE, SARIMA, SARIMA-ANN, and SEP.

INTRODUCTION

Agricultural prices forecasts are intended to be useful for farmers, governments and agribusiness industries. The ability to accurately forecast the agricultural commodities price therefore an important concern in both policy maker and business circles. Time series approaches are commonly used to forecast prices. The Autoregressive Integrated Moving Average (ARIMA) model is a popular time series model used to forecasting purpose. The ARIMA model's popularity stems from its statistical properties as well as the well-known Box-Jenkins methodology used in model construction. However, the main disadvantage of the ARIMA model is that it cannot capture nonlinear behavior. Although linearity is a beneficial assumption and an effective tool in many aspects, it became straightforward in the early 1980s that estimating complex real-world problems with linear models is not always appropriate. Over the last two decades, a number of studies on nonlinearity of time series data has increasing rapidly.

The SARIMA model is a time series model used to deal with the seasonal and linear behavior of agricultural commodity price and arrival series. Kumar *et al.* (2011) used SARIMA model to forecast future tomato prices for the rainy season harvest period, August to October. Model identification and estimation for forecasting tomato prices using the SARIMA model were performed only after the price series were transformed to make them stationary. Adanacioglu and Yercan (2012) investigated seasonal tomato price variation and developed a SARIMA model to forecast monthly tomato prices at the wholesale level based on observed prices over a ten-year period. Keerthi and Naidu (2013) forecast monthly tomato prices using a univariate ARIMA model fitted with historical data. This model is useful when time series exhibit nonlinear dynamic behavior such as asymmetry, frequency amplitude dependence, and volatility clustering. In this paper, the wholesale price and arrival series exhibit a combination of linear and nonlinear behavior, prompting the use of a hybrid model for predicting agricultural commodity prices and arrivals. In general, time series data exhibit both linear and nonlinear behavior; no single model can identify all of the characteristics of time series data. Yollanda and Devianto (2020) developed a hybrid SARIMA-ANN model to forecast tourist arrivals at Minangkabau International Airport. Udayshankar and Sharma (2020) identified the SARIMA$(1,0,0)(2,1,0)_{12}$ model as an adequate and suitable model for forecasting egg prices in Telangana. Gangshetty *et al.* (2021) used SARIMA (Seasonal Autoregressive Integrated Moving Average) model for temperature prediction of Pune, Maharashtra. Wu *et al.* (2021) proposed a novel hybrid approach SARIMA + LSTM (seasonal autoregressive integrated moving average combined with long short-

term memory) to predicted daily tourist arrivals to Macau SAR, China. The primary goal of this paper is to compare the forecasting performance of the SARIMA and SARIMA-ANN models. Kumari *et al.* (2022) compared the Statistical Models for Prediction Area, Production and Yield of Citrus in Gujarat.

RESEARCH METHODOLOGY

The monthly wholesale price and arrival time series of tomato crop in Gurugram market in Haryana from January 2010 to December 2021 were used in the study. The first 132 observations (from January 2010 to December 2020) are used for model building and parameter estimation, while the next 12 observations (from January 2021 to December 2021) are used for post-validity checking. The dataset for the analysis was obtained from the website https://agmarknet.gov.in.

SARIMA model. SARIMA is a commonly used model for linear and seasonal time series analysis and forecasting. A time series has N total number of observations that is denoted by $\{y_t|t = 1, 2, ..., N\}$ is created by a SARIMA (p, d, q)× (P, D, Q)s process by following equation:

$$\emptyset(B)(1-B)^d\phi(B^s)(1-B^s)^D y_t = c + \theta(B)\Theta(B)\varepsilon_t$$

Where $\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 \dots\dots \emptyset_p B^p$

$$\phi(B^s) = 1 - \phi_1 B^s - \phi_2 B^{2s} \dots\dots - \phi_P B^P$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 \dots\dots - \theta_p B^p$$

$$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} \dots\dots - \Theta_P B^P$$

$$SARIMA = ARIMA \underbrace{(p,d,q)}_{Non-seasonal\ part} \times \underbrace{(P,D,Q)_s}_{Seasonal\ part}$$

B is back shift operator, s is number of observations in a year, $\emptyset(B)$is the non-seasonal Auto-Regressive operator (AR) of order p, $\phi(B)$ is the Seasonal Auto-Regressive operator (SAR) of order P, $\theta(B)$ is the non-seasonal Moving Average operator (MA) of order q, $\Theta(B)$ is of the Seasonal Moving Average operator (SMA) with order Q, d is order of non-seasonal differencing.

– D is order of seasonal differencing.

$$\begin{cases} D = 0, \ time\ series\ has\ no\ seasonality\ effect \\ D \geq 1, time\ series\ has\ seasonality\ effect \end{cases}$$

– ε_t is considered residual.

Hybrid (SARIMA-ANN) model. Hybrid model is a combination of linear and nonlinear models that is typically used to improve forecast accuracy. In general, the mathematical form of combining linear and nonlinear models is as follows:

$$y_t = L_t + N_t + \varepsilon_t$$

Where L_t is a linear component and N_t is a nonlinear component of the model. In this paper, the SARIMA model is used in the first step to handle the linear component. Assume thee$_t$ (residual) is a nonlinear component obtained from the SARIMA model.

$$e_t = Y_t - \widehat{L_t}$$

Where $\widehat{L_t}$is estimated value from linear SARIMA model at period t. Then, the second step, ANN model is used for modelling residual series from SARIMA at as follows:

$$e_t = f(e_{t-1}, e_{t-2} \dots\dots e_{t-p}) + e_t = \widehat{N_t} + \varepsilon_t$$

Where $\widehat{N_t}$ is estimated value and ε_t is the residual of non-linear ANN model at period t. Hence, the forecast value of the hybrid ARIMA-ANN model is as follows:

$$\widehat{Y_t} = \widehat{L_t} + \widehat{e_t}$$

Models Performance Statistics. In this paper, two main statistics, including Mean Absolute Error (MAE) and percent standard error of prediction (SEP), are proposed to evaluate the forecasting performance of the SARIMA and SARIMA-ANN models, which are formulated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$SEP = \frac{100}{\bar{y}} RMSE, \quad where\ RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_t)^2}$$

Where y_t and $\hat{y}_t$ are the actual and predicted values for time t. n is the number of observations in validity set.

RESULT AND DISCUSSION

There is a trend and seasonality in the tomato price and arrivals series, so both are non-stationary. Stationarity is a required condition for using the SARIMA model. The observed series is decomposed into three components: trend, seasonal, and random. By removing the estimated trend and seasonal components from the actual time series, the estimated random component is obtained. By removing the estimated seasonal component from observed price and arrival series, de-seasonal time series are obtained Gope *et al.* (2022). After one order non seasonal difference of de-seasonal time series, the stationary series is obtained. Fig. 1 and 2 show ACF/PACF plots of non-stationary and stationary price and arrival series. The ACF plot of a stationary series shows that the spikes are outside the 2 limits at lag 2 and 12. It indicates that the maximum possible order of q and Q for price series is 2 and 1, respectively. PACF for stationary series reveals that spikes outside the 2 limits occur at lags 2, 12, and 24, implying that the maximum possible order for p and P order is two.

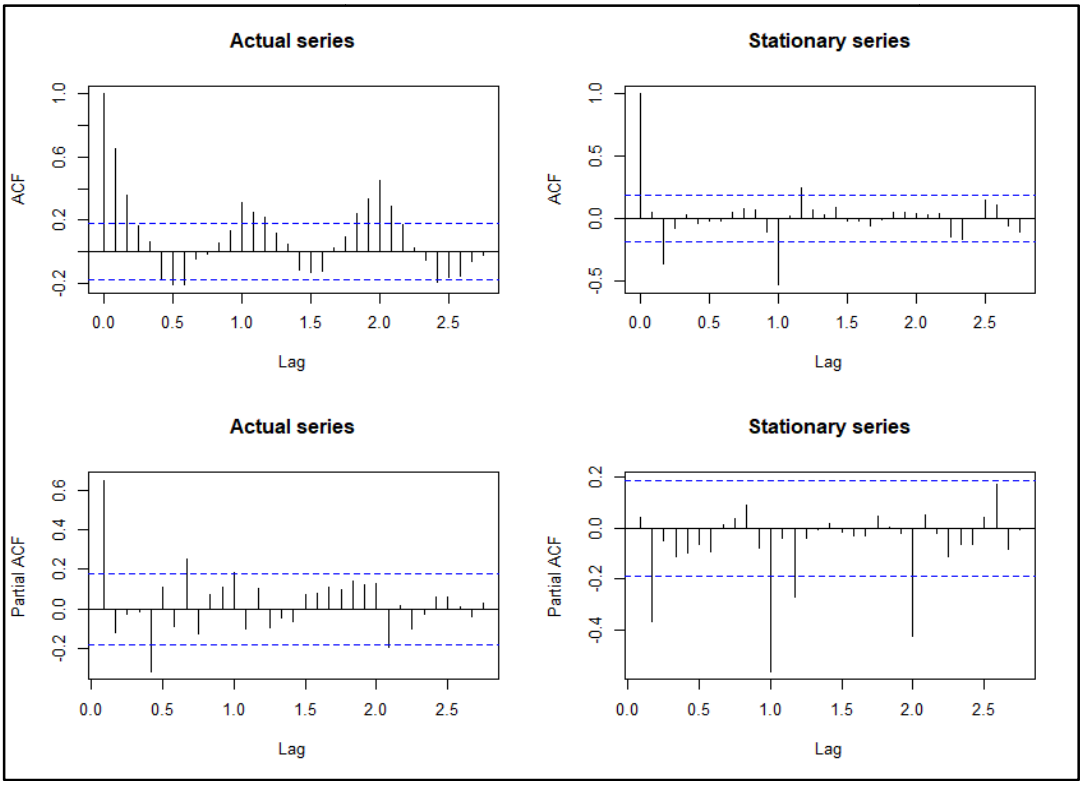

Fig. 1. ACF/PACF plots of price series.

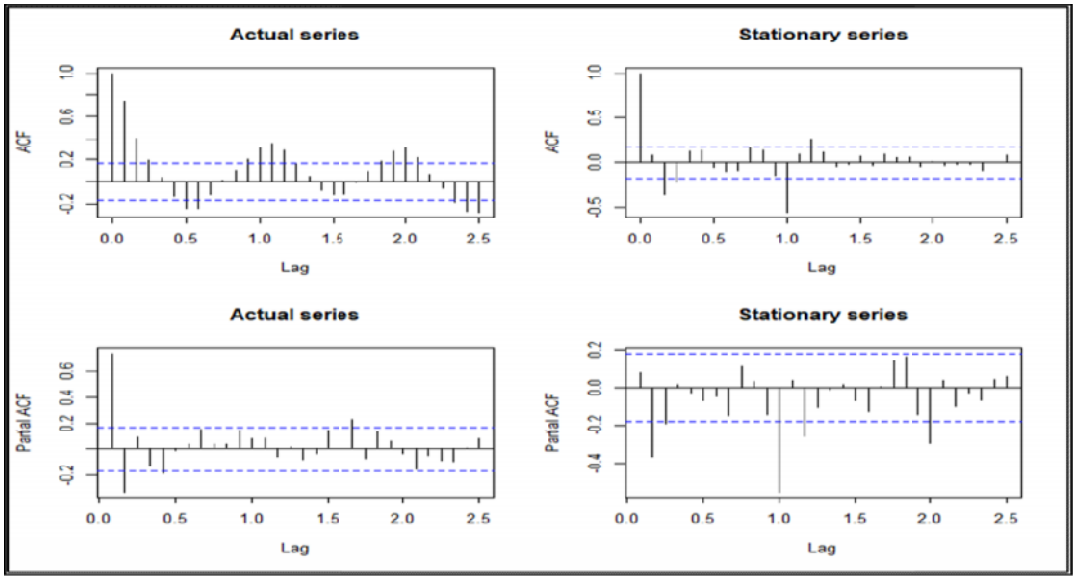

Fig. 2. ACF/PACF plots of arrival series.

Next step, to analyze the goodness of fit, three measures of goodness are used: the Akaike information criterion (AIC), the Root mean squared error (RMSE), and the Mean absolute percentage error (MAPE). Table 1 shows the values of these measures.

Table 1: SARIMA Model Goodness of Fit Measures.

SARIMA	AIC	RMSE	MAPE
Prices			
$(0,1,2)(0,1,1)_{12}$	1649.15	430.93	21.32
$(2,1,0)(1,1,1)_{12}$	1652.55	432.58	19.04
$(1,1,1)(1,1,1)_{12}$	1655.17	437.45	28.83
$(2,1,1)(0,1,1)_{12}$	1654.53	432.71	28.97
$(1,1,2)(0,1,1)_{12}$	1656.82	429.11	28.56
$(1,1,2)(1,1,1)_{12}$	**1644.38**	**410.51**	**17.21**
Arrivals			
$(1,1,1)(0,1,1)_{12}$	1714.02	1152.41	36.16
$(1,1,1)(1,1,0)_{12}$	1712.97	1171.57	37.69
$(1,1,0)(1,1,0)_{12}$	1713.3	1159.45	35.97
$(1,1,0)(0,1,1)_{12}$	1823.35	1259.29	37.29
$(1,1,0)(0,1,0)_{12}$	1802.94	1271.41	37.75
$(2,1,2)(0,1,0)_{12}$	**1710.79**	**1132.28**	**34.05**
$(1,1,1)(0,1,1)_{12}$	1814.02	1152.41	38.16

Table 2: Estimated parameters of selected SARIMA model for training data set.

	Estimate	SE	Z value	P- value
Prices				
AR1	0.29	0.13	2.23	0.02
MA1	-0.34	0.12	-3.56	<0.01
MA2	-0.57	0.11	-5.03	<0.01
SAR1	-0.22	0.09	-2.19	0.02
SMA1	0.97	0.21	-4.71	<0.01
Arrivals				
AR1	-0.36	0.14	-2.47	0.01
AR2	-0.51	0.07	-12.85	<0.01
MA1	0.21	0.08	1.53	0.01
MA2	0.70	0.11	7.19	<0.01

$SARIMA(1,1,2)(1,1,1)_{12}$ and $SARIMA(2,1,2)(0,1,0)_{12}$ are selected for residual analysis as the best among the examined models with the lowest AIC, RMSE, and MAPE values and significance parameter estimation for tomato prices and arrivals series in Gurugram market. Fig. 3 and 4 show that the Ljung-Box statistic is non-significant, indicating that there is no autocorrelation in the residual series. As a result, these models are appropriate for forecasting prices and arrivals in the Gurugram market.

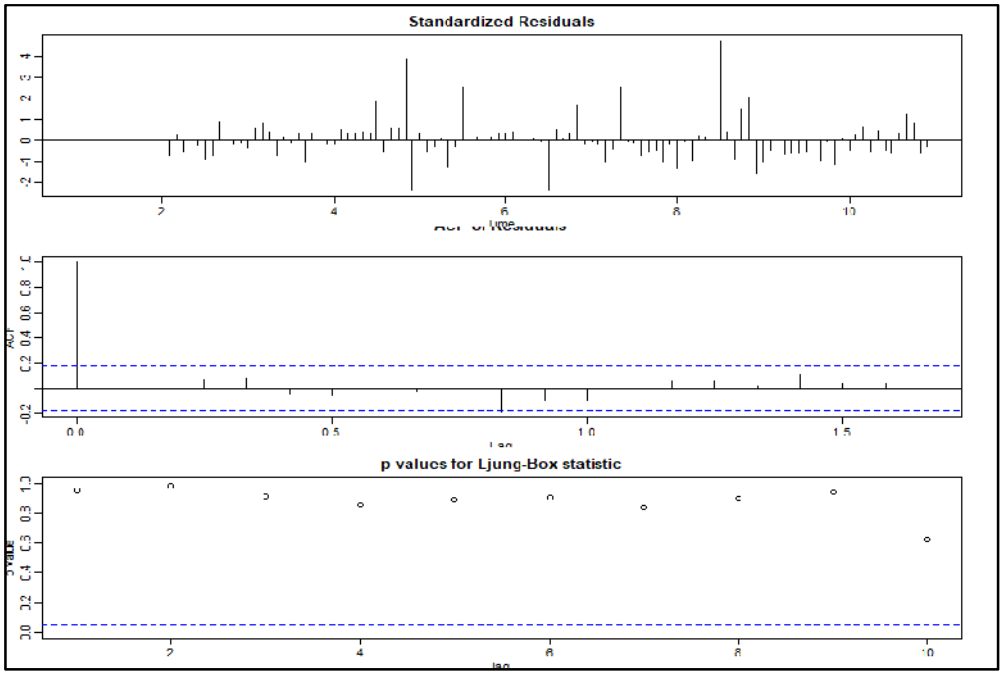

Fig. 3. Plot of Residuals from SARIMA Model for Price in Gurugram Market.

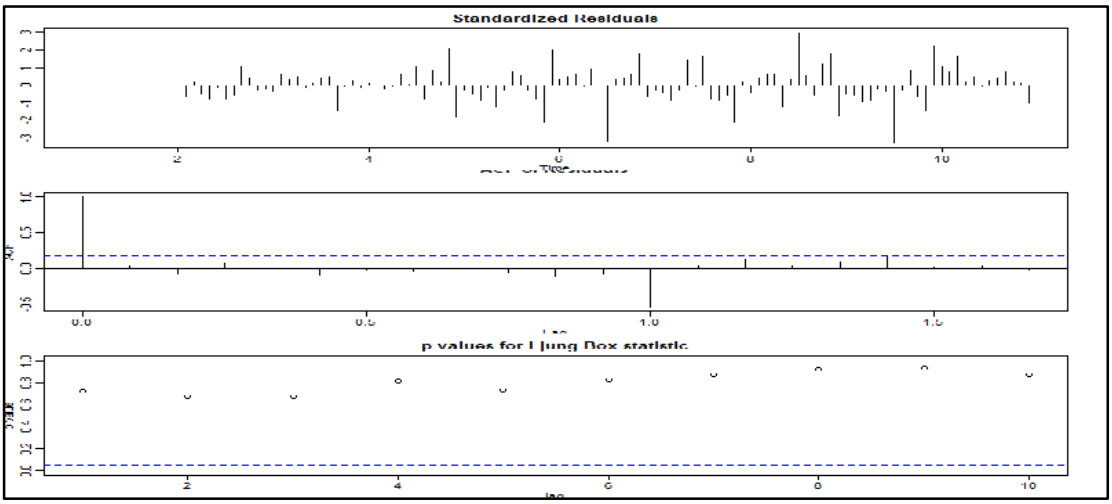

Fig. 4. Plot of Residuals from SARIMA Model for arrivals in Gurugram Market.

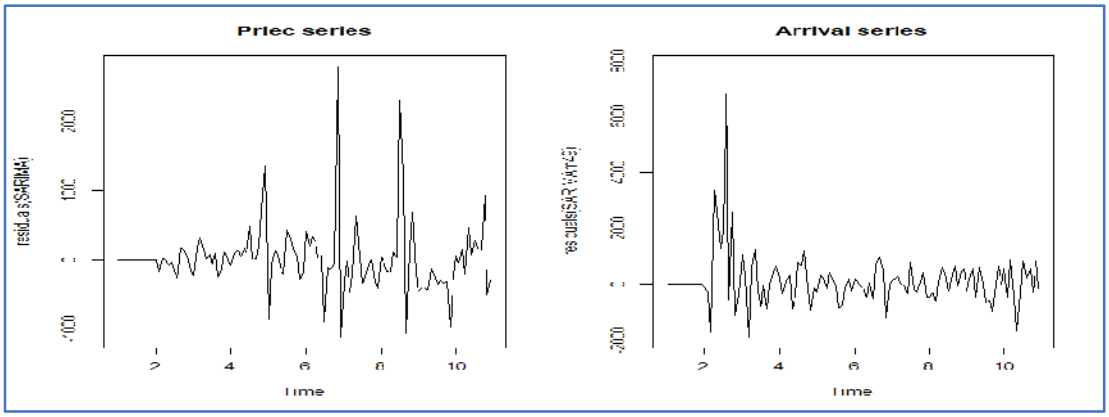

Fig. 5. Residual plots of fitted SARIMA models.

The residuals from the fitted model provide a clear hint for improvement. Fig. 5 depicts residuals from selected models, which indicate nonlinearity. Because traces of nonlinearity behavior are discovered, an artificial neural network (ANN) model is used to model the residuals obtained from SARIMA. The ANN model used here is a three-layered feed forward neural network that was trained for 10x106 epochs with a learning rate of 0.001 and a threshold of 0.01. Several practical considerations were used to determine the optimal number of neurons in the hidden layer. The number of nodes in the hidden layer is optimized using a trial-and-error method. The hidden layer with different neurons and activation functions was used to find the best fitted model. The networks with the lowest RMSE (Root mean squared error) and MAPE (mean absolute percentage error) in the testing set are chosen as the best ANN for the corresponding series. Table 3 displays the RMSE and MAPE values for the ANN model. Table 4 displays predicted values for price and arrival based on the best-fit SARIMA and SARIMA-ANN models.

Table 3: RMSE and MAPE values for the ANN model for residuals series of SARIMA.

Input Layer	Number of Artificial neurons	Act. Fct.	RMSE	MAPE
Prices				
12	4	Tanh	408.40	18.73
12	**5**	**Tanh**	**390.30**	**17.83**
12	8	Tanh	405.27	19.70
12	9	Logistic	495.71	22.82
12	10	Logistic	491.71	22.10
12	11	Logistic	498.40	20.73
Arrivals				
12	3	Tanh	685.10	82.83
12	**4**	**Tanh**	**621.37**	**80.68**
12	6	Tanh	681.60	89.06
12	8	Logistic	715.10	99.83
12	9	Logistic	719.37	95.68
12	10	Logistic	751.60	97.06

Table 4: Observed and predicted values of SARIMA and Hybrid model.

Month	Observed Price	SARIMA Predicted	Hybrid Predicted	Observed Arrival	SARIMA Predicted	Hybrid Predicted
Jan-21	800.43	921.13	1016.855	1008	1167.92	1295.67
Feb-21	800.43	1102.89	1121.822	1008	1146.30	1199.07
March-21	804.35	1037.65	1148	775.8	1012.05	1226.33
April-21	804.35	1151.14	954.842	563.6	1067.33	854.96
May-21	509.86	882.74	639.347	837.2	1155.78	953.71
June-21	509.86	890.07	701.751	837.2	1523.26	1098.22
July-21	1410.4	1524.52	1310.764	708	1235.22	1021.45
Aug-21	1338.79	1410.57	1533.609	708	1134.51	932.62
Sept-21	3579.17	1883.37	2971.556	869.6	1274.00	1097.26
Oct-21	3579.17	2757.48	2873.821	869.6	1079.98	813.25
Nov-21	2811.88	2105.72	2354.068	1181.6	1365.12	1231.11
Dec-21	2070.22	2022.53	2125.765	1181.6	1581.29	1427.58

Table 5 displays the performance statistics of the SARIMA and Hybrid (SRIMA-ANN) models for the validation data set with various forecast horizons of price and arrival series in terms of SEP ratio and MAE ratio. SEP and MAE ratios of both models, calculated as SEP (SARIMA)/SEP (SARIMA-ANN) and MAE (SARIMA)/MAE (SARIMA-ANN). If the SEP and MAE ratios are greater than one, the Hybrid SARIMA-ANN model outperforms the SARIMA model. Table 5 shows that SEP and MAE ratios greater than one for 6, 9, and 12 months ahead forecast indicate that the hybrid SARIMA-ANN model performed better, but SARIMA model performed better for 1 and 3 months ahead forecast for both series.

Table 5: Comparative results in terms of SEP and MAE ratios between SARIMA and Hybrid model.

Series	1 Month	3 Months	6 Months	9 Months	12 Months
SEP ratio					
Prices	0.54	0.77	1.27	2.13	1.77
Arrivals	0.56	0.55	1.37	1.49	1.53
MAE ratio					
Prices	0.55	0.74	1.29	1.61	1.50
Arrivals	0.56	0.57	1.27	1.43	1.54

SUMMARY AND CONCLUSION

This study compared the modelling and forecasting performance of SARIMA and Hybrid (SARIMA-ANN) models using monthly wholesale price and arrival series of tomato crops in Gurugram market of Haryana. The goal of this study is short term forecast up to one year with different forecast horizons, such as 1, 3, 6, 9 and 12 months. SARIMA$(1,1,2)(1,1,1)_{12}$ and SARIMA$(2,1,2)(0,1,0)_{12}$ are the suitable models for capturing the linear pattern of price and arrival series, with the lowest AIC, RMSE, and MAPE values and significance parameter estimation. In comparison to the SARIMA and Hybrid models, the hybrid models provide better forecasting accuracy in terms of the lowest value of performance statistics such as MAE and SEP for 6, 9, and 12 months ahead forecast, whereas the SARIMA model performs better for 1 and 3 months ahead forecast.

FUTURE SCOPE

We found a significant fluctuation in the data in this study, implying that nonlinear models like GARCH and ANN may be used to improve the forecasting performance.

Acknowledgements. The authors are thankful Department of mathematics and Statistics in CCS HAU, Hisar for providing the research facilities for this research. Special thanks to UGC-NFSC for their financial support, which assisted my research to be successful.

Conflict of Interest. None.

REFERENCES

Adanacioglu, H. and M. Yercan, M. (2012). An analysis of tomato prices at wholesale level in Turkey: An application of SARIMA model. *Custos e Agronegocio, 8*, 52-75.

Gangshetty, A., Kaur, G. and Maluje, U. S. (2021). Time series prediction of temperature in Pune using Seasonal ARIMA model. *International Journal of Engineering Research & Technology*, *10*(11), 235-240.

Gope, P. Selvi, R. P., Vasanthi, R. and Karthick, V. (2022). A Study on Price Forecasting of Paddy in West Tripura, District. *Biological – An Forum International Journal*, *14*(2), 1469-1473.

Keerthi, P. K. and Naidu, G. M. (2013). Forecasting monthly prices of tomato in Madanapalli market of Chittoor district. *Bioinfolet*, *10*(1b), 201-203.

Kumar, T. L. M., Surendra, H. S. and Munirajappa, R. (2011.) Holt-winters exponential smoothing and seasonal ARIMA time-series technique for forecasting of onion price in Bangalore market. *Mysore journal of agricultural sciences*, *45*(3), 602-607.

Kumari, P., Parmar, D. J., Mahera, A. B. and Lad, Y. A. (2022). Comparison of Statistical Models for Prediction Area, Production and Yield of Citrus in Gujarat. *Biological Forum – An International Journal*, *14*(2), 690-695.

Udayshankar, M. and Sharma, M. R. (2020). Forecasting of egg prices in Telangana using R-Software. *Indian Journal of Applied Research*, *10*(3), 44-46.

Wu, D. C. W., He, L. J. L. and Tso, K. F. G. (2021). Forecasting tourist daily arrivals with a hybrid SARIMA– LSTM approach. *Journal of Hospitality & Tourism Research*, *45*(1), 52-67.

Yollanda, M. and Devianto, D. (2020). Hybrid model of seasonal ARIMA-ANN to forecast tourist arrivals through Minangkabau international airport. *Department of Mahematics, Andalas University, Padang*, 3(5), 755-765.

How to cite this article: Pushpa, Joginder Kumar and Vikram (2022). Hybrid Sarima-Ann Model for Forecasting Monthly Wholesale Price and Arrival Series of Tomato Crop. *Biological Forum – An International Journal*, *14*(4a): 591-596.