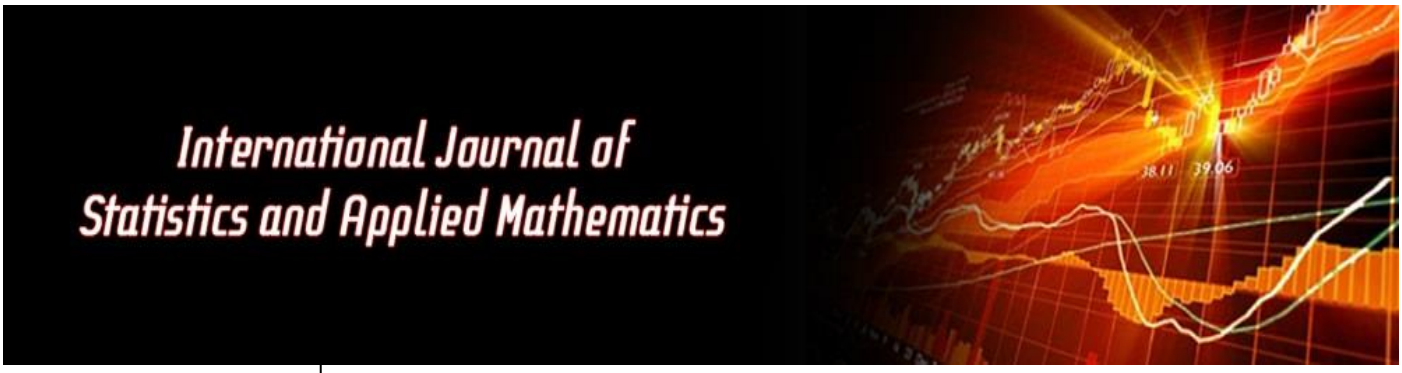
International Journal of Statistics and Applied Mathematics 2024; 9(2): 101-107



ISSN: 2456-1452
Maths 2024; 9(2): 101-107

www.mathsjournal.com
Received: 06-01-2024
Accepted: 08-02-2024

Pushpa Ghiyal
Department of Mathematics and Statistics, CCSHAU, Hisar, Haryana, India

Joginder Kumar
Department of Mathematics and Statistics, CCSHAU, Hisar, Haryana, India

Corresponding Author:
Pushpa Ghiyal
Department of Mathematics and Statistics, CCSHAU, Hisar, Haryana, India

Use of hybrid SARIMA-GARCH model for predicting the prices of agricultural product in Haryana

Pushpa Ghiyal and Joginder Kumar

Abstract
In this paper, Seasonal Autoregressive Integrated Moving Average (SARIMA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and Hybrid (SARIMA-GARCH) models have been used for modelling the prices of tomato in Panipat APMC (Agricultural Produce Market Committee) market of Haryana. Akaike information criteria (AIC) and Bayesian information criteria (BIC) have been used as model selection criteria. And, forecasting performance measure such as relative percentage deviation (RD (%)), mean absolute deviation error (MAPE) and standard error of prediction (SEP) have also been used to check the accuracy of the fitted models. The results of present study showed that the performance of Hybrid models was more appropriate as compared to SARIMA and GARCH models for predicting the price of agricultural product (tomato). The findings of the present study will help in taking decisions in managing agricultural supplies.

Keywords: SARIMA, GARCH, Hybrid, APMC, AIC and BIC

Introduction
Time series forecasting is an important area in which prediction of future values of a variable is made based on past values of the variable. Slutsky (1937) [11] and Yaglom (1955) [12] first formulated the concept of Autoregressive (AR) and Moving Average (MA) models. Box and Jenkins (1970) [2] integrated the existing knowledge in the book entitled "Time Series Analysis: Forecasting and Control" which has an enormous impact on the theory and practice of modern time series analysis and forecasting. Chandran and Pandey (2007) [3] forecasted the prices of potatoes for Delhi market using univariate seasonal model ARIMA (1, 1, 1), (1, 0, 0) and found that Short-term forecasts based on this model were close to the observed values. Hakan and Murat (2012) [5] studied seasonal price variation of tomato crops and created Seasonal ARIMA (SARIMA) model to forecast monthly tomato prices at the wholesale level for Antalya in Turkey. In that study SARIMA (1, 0, 0) (1, 1, 1) model was selected as the best model to forecast tomato prices.
Modelling and forecasting of volatility by nonlinear models has emerged as an important tool for time-series analysis. For this, the most commonly used statistical models are Autoregressive Conditional Heteroscedastic (ARCH) models (Engle 1982) [4] and Generalized ARCH (GARCH) models (Bollerslev 1986) [1]. Jordaan *et al.* (2007) studied the volatility in the prices of white maize, yellow maize & sunflower seed using GARCH approach. The volatility in the prices of wheat and soybeans was found constant over time.
Agriculture time series data usually contain both linear and nonlinear patterns. Therefore, no single (linear or nonlinear) model can be adequate to identify all the characteristics of time series data as linear (SARIMA) model cannot capture nonlinear pattern while nonlinear (GARCH) model cannot capture linear pattern. A hybrid time series models combine linear and nonlinear models to capture different patterns in time series data and to improve forecasting performance. SARIMA-GARCH hybrid models have been used in this study. Malik (2015) presented the modelling and forecasting performance of ARIMA, GARCH (1,1) and mixed ARIMA - GARCH (1,1) models using historical daily close price data from the NASDAQ stock exchange for GE company (USA). Shetty *et al.* (2018) [10] proposed the hybrid model of the linear seasonal autoregressive moving average (SARIMA) and the

non-linear generalized autoregressive conditional heteroscedasticity (GARCH) for forecasting the gold price. The goodness of fit of the model was measured AIC, while the performance measure was assessed by using RMSE, MAE and MAPE and concluded that SARIMA-GARCH was more appropriate model for forecasting the gold price. Mallikarjuna *et al.* (2019) [9] studied the forecasting performance of time-series models such as ARIMA and ARCH models for the prices of black pepper in one of the major markets of Karnataka state. Yollanda and Devianto (2020) [13] built SARIMA-ANN hybrid model to forecast tourist arrivals through Minangkabau International Airport.

Methodology

Seasonal Autoregressive Integrated Moving Average Model

Seasonal Autoregressive Integrated Moving Average (SARIMA) is a time series linear technique, based on traditional ARIMA technique, widely used for modelling of seasonality. In ARIMA (p, d, q) technique, future value of a variable is a linear function of several past observations and random errors. There are six parameters for fitting the Seasonal ARIMA (p, d, q), models: Order of autoregressive (p) and seasonal autoregressive (P), order of integration (d) and seasonal integration (D) & order of moving average (q) and seasonal moving average (Q) and s represents the season period length. The general expression of the model is given below.

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

Autocorrelation Function (ACF)

The autocorrelation function is most important tools for dependence in a series. If stationarity is assumed and autocorrelation function ρ_k for a set of lags K = 1, 2. can be estimated by simply computing the sample correlation coefficient between the pairs, k units apart in time. The correlation coefficient between y_t and $y_{t-k}$is known as autocorrelation or serial correlation coefficient of $y_t$and represented by the symbol ρ_k, which is defined as.

$$\rho_k = \frac{\sum_{t=1}^{n-k}(y_t-\overline{y})(y_{t+k}-\overline{y})}{\sum_{t=1}^{n}(y_t-\overline{y})^2} \ldots 2$$

It ranges from -1 to +1. The maximum number of ρ_k is approximately N/4, where N is the number of y_t observations.

Partial Autocorrelation Function (PACF)

The correlation coefficient between two random variables y_t and y_{t-k} after removing the effects of the intervening variables $y_{t-1}, y_{t-2}, \ldots, y_{t-t-k+1}$ is known as partial autocorrelation and represented by ϕ_{kk}, which is defined as:

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1}\emptyset_{k-1,j}\rho_{k-j}}{1-\sum_{j=1}^{k-1}\emptyset_{k-1,j}\rho_{k-j}} \ldots 3$$

Ljung Box Test

Ljung Box (1978) proposed a test statistic that is based on all residual autocorrelations with the following hypotheses:
H_0: the residuals are independently distributed
H_1: the residuals are not independently distributed
The test statistic is

$$Q = n(n+2)\sum_{k=1}^{P}(n-k)^{-1}r_k^2(\hat{a}_t) \quad \ldots 4$$

The statistic Q follows a Chi-squared distribution with (p-m) degrees of freedom where n is the total number of observations used to estimate the model, p is the number of residual autocorrelations and m is the number of estimated parameters of the model. The model is considered appropriate if Q statistic is significant at 5% level of significance.

Generalized Autoregressive Conditional Heteroscedasticity Model

The generalized autoregressive conditional heteroskedasticity (GARCH) model is widely used to model the volatility in time series data. Bollerslev (1986) [1] proposed the GARCH model as a generalization of ARCH model, which allows conditional variance to be dependent on previous own lags in the same way that in ARMA process.
The GARCH (p, q) model for the series ε_t is given by

Let AR (p) mean model

$$y_t = \theta_0 + \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} + \varepsilon_t \quad \ldots 5$$

$$\varepsilon_t = \sigma_t^2 \xi_t$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q}\alpha_i\varepsilon_{t-i}^2 + \sum_{i=1}^{p}\beta_i\sigma_{t-i}^2 \quad \ldots 6$$

Jarque-Bera (JB) Test

The Jarque-Bera (1980) [6] test determines whether sample data have skewness and kurtosis that match a normal distribution. J.B. statistic is defined as:

$$J.B. = \frac{N}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \quad \ldots 7$$

The Jarque-Bera statistic is distributed as chi-square with 2 degrees of freedom with the following hypotheses:
H_0: Time series is normally distributed
H_1: Time series is not normally distributed

Hybrid Models

Time series models can also be used as a composition of linear autocorrelation structure and a non-linear component which is given in mathematical form.

$$y_t = L_t + N_t \quad \ldots 8$$

Where, $y_t$is the observation at time tand L_t&N_t denote linear and nonlinear components, respectively. These two components can be estimated from the data. SARIMA model is fitted to the linear component and the corresponding forecast $\hat{L}_t$at time tis obtained. So, the residual at time *t* is given by.

$$N_t = e_t = y_t - \hat{L}_t \quad \ldots 9$$

The residuals dataset after fitting SARIMA model contains only nonlinear component and that can be properly modelled through an GARCH model.

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \; or \; \hat{y}_t = \hat{L}_t + \hat{e}_t \quad \ldots 10$$

Comparison and Validation of the Developed Models

Model Selection

Information criteria such as AIC and BIC are used to select an appropriate model.

$$\text{AIC} = -2In(\text{L}) + 2k \quad \ldots 11$$

$$\text{BIC} = -2In(\text{L}) + In(\text{N})k \quad \ldots 12$$

Where L is the value of the likelihood function evaluated at the parameter estimates, N is the number of observations and k is the number of estimated parameters.

Model Evaluations

The model is evaluated quantitatively MAPE, SEP) (%), RMSE and RD%. The SEP is used for the comparison of forecast from different models because of its dimension less.

$$\text{MAPE} = \frac{100}{\text{n}} \times \sum_{\text{i}=1}^{\text{n}} \left|\frac{\text{O}_\text{i} - \text{E}_\text{i}}{\text{O}_\text{i}}\right| \quad \ldots 13$$

$$\text{SEP} = \frac{100}{\bar{\text{y}}}\text{RMSE where RMSE} = \left[\frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{n}}(\text{O}_\text{i} - \text{E}_\text{i})^2\right]^{\frac{1}{2}} \quad \ldots 14$$

$$RD = \frac{O_i - E_i}{O_i} \times 100 \quad \ldots 15$$

Where, O_i, $\bar{y}$ and E_i are the observed, mean and predicted values and n is the number of observations in validation set.

Results

Seasonal Autoregressive Integrated Moving Average (SARIMA) model

From Figures 1, it can be seen that the order of MA (q) and SMA (Q) is expected to be lie between $0 \leq q \leq 2$ and $0 \leq Q \leq 1$ and order of AR (p) and SAR (P) is expected to be lie between $0 \leq p \leq 2$ and $0 \leq P \leq 1$ for prices of tomato in Panipat market. From Figures 4.18, it can be seen that the order of MA (q) and SMA (Q) is expected to be lie between $0 \leq q \leq 3$ and $0 \leq Q \leq 1$ and order of AR (p) and SAR (P) is expected to be lie between $0 \leq p \leq 2$ and $0 \leq P \leq 2$ for arrivals of tomato in Panipat market.

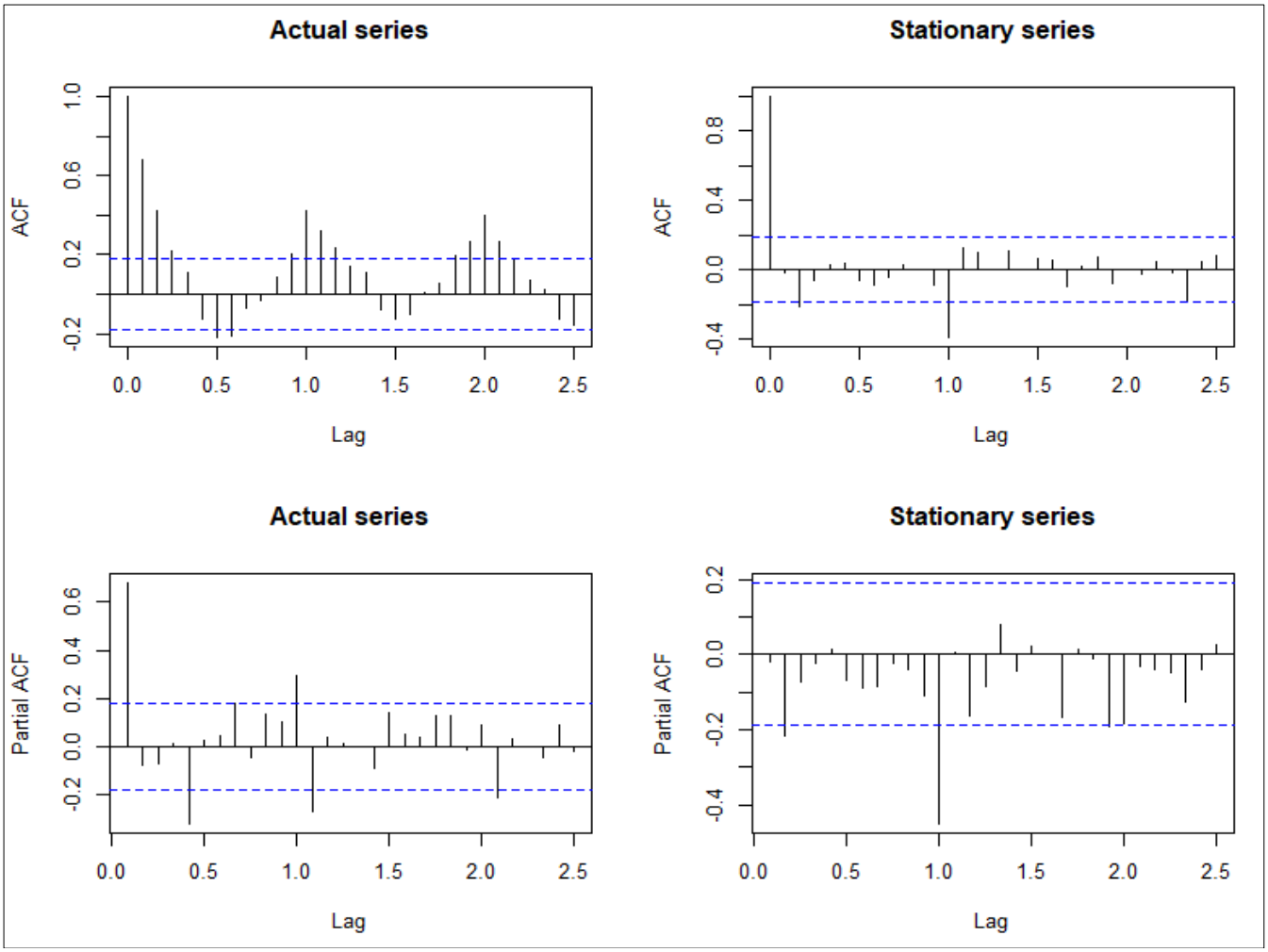


Fig 1: ACF and PACF Plots of actual and stationary series

Table 1: Selection criteria of SARIMA models

SARIMA	Criteria		
	AIC	RMSE	MAPE
$(1,1,1)(1,1,1)_{12}$	1657.36	434.44	19.30
$(1,1,1)(0,1,1)_{12}$	1655.65	432.34	19.13
$(2,1,1)(0,1,1)_{12}$	1655.47	428.93	18.54
$(2,1,1)(1,1,1)_{12}$	1657.24	429.32	18.55
$(1,1,2)(0,1,1)_{12}$	1654.68	427.87	17.52
$(2,1,2)(0,1,1)_{12}$	1656.25	428.13	18.47

Table 2: Estimated parameters of SARIMA (1, 1, 2), $(0, 1, 1)_{12}$ model

Estimate		S.E.	z-value	p-value
AR1	0.52	0.14	3.77	<0.01
MA1	-0.71	0.23	-3.07	<0.01
MA2	-0.27	0.14	-1.85	0.04
SMA1	0.98	0.50	-1.96	0.04
Ljung-Box	Statistic	0.12	p-value	0.71

Table 2 shows that the estimated parameters of the selected model for prices (AR1, MA1, MA2, SMA1) and arrivals (MA1, SMA1) of tomato in Panipat market, are found significant at 5% level of significance. It can also be observed that the coefficients of all parameters meet the condition of stationary and in vertibility for SARIMA model. The values of Ljung-Box "Q" statistic for all selected models are found non-significant as p-value is greater than 0.05 which indicating residuals have white noise (no autocorrelation). Thus, on the basis of above results, it is observed that the SARIMA (1, 1, 2), $(0, 1, 1)_{12}$ model was appropriate price of tomato in Panipat market.

Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model

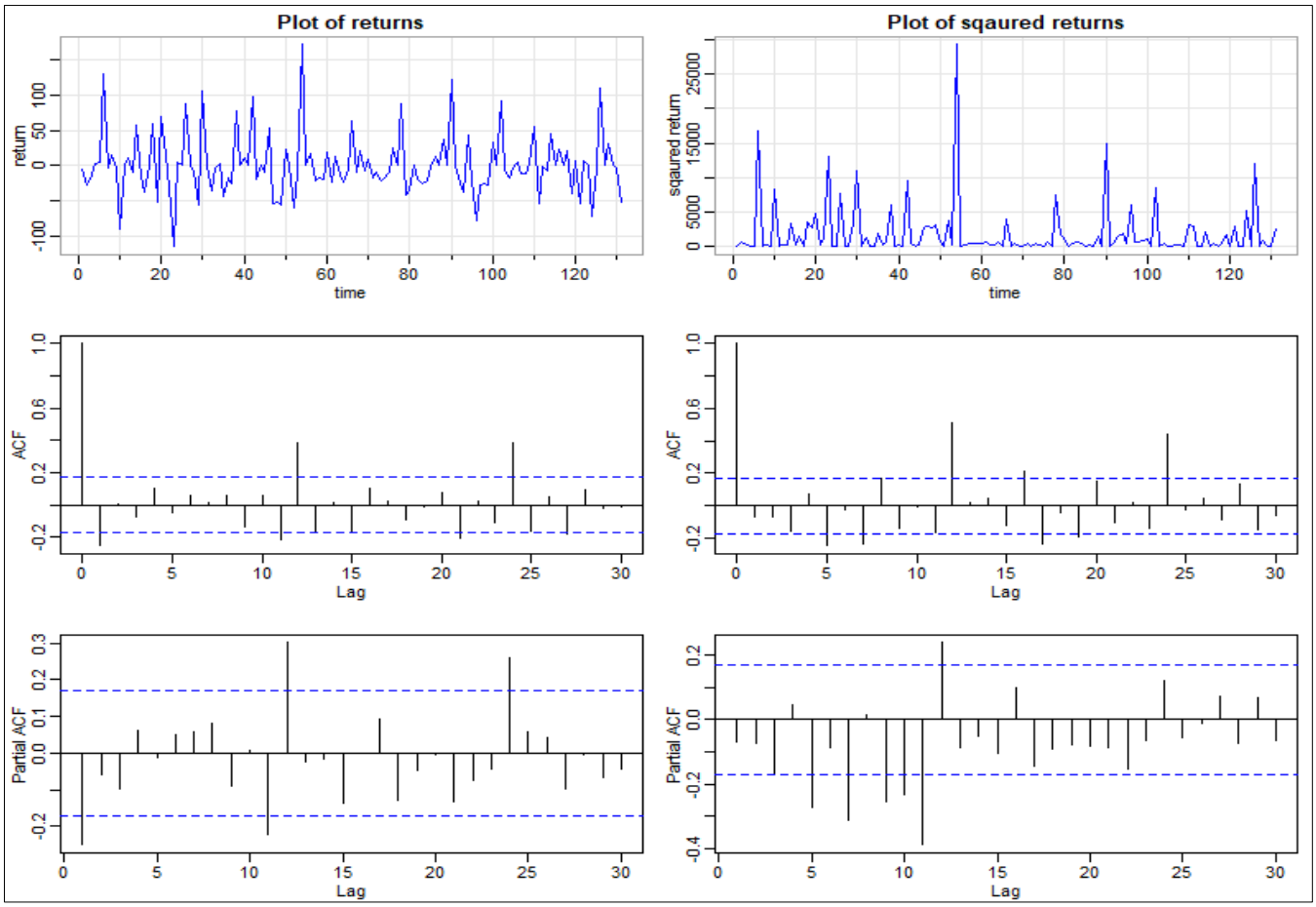


Fig 2: Plots of returns and squared returns

Figure 2 reveals that returns are not independent and volatility clustering is apparent. From, ACF and PACF plots of returns revealed that ACF is significant at lag 2, 11, 12, 21 & 24 and PACF is significant at lag 1, 11, 12 & 24. The ACF and PACF plots of squared returns are revealed that ACF is significant at lag 5, 7, 11, 16, 17 & 24 and PACF is significant at lag 5, 7, 9, 10, 11 & 12. Hence, the returns and squared returns for prices of tomato in Panipat market are auto-correlated. As a result, GARCH model is applicable to this series.

Table 3: Estimated parameters of ARMA (1,1)-GARCH (0,1) model for returns series

Parameter	Estimate	S.E.	t-value	p-value
μ	0.72	0.12	5.63	<0.01
AR1 ($\emptyset_1$)	0.70	0.06	10.82	<0.01
MA1 (θ_1)	-0.99	0.01	-156.95	<0.01
α_0	0.01	0.01	0.01	<0.01
β_1	0.96	0.02	383.47	<0.01
Skew	1.52	0.16	9.07	<0.01
Shape	5.99	2.53	2.36	<0.01
Model selection criteria	AIC	15.56	BIC	15.68
Diagnostic checking	Ljung-Box		ARCH LM	
	Statistic	p-value	Statistic	p-value
	0.67	0.92	0.76	0.96

The results of Ljung-Box and ARCH-LM tests are given for standardized residuals obtained from the selected ARMA (1, 1) - GARCH (0, 1) model for returns of price of tomato. It can be observed that the results of Ljung-Box and ARCH-LM tests are not significant at 5% level of significance as p-value for both statistic(s) is greater than 0.05. Hence, the null hypothesis is not rejected and it is concluded that there is no autocorrelation and no ARCH effect in standardized residuals. On the basis of results so obtained, ARMA (1, 1) - GARCH (0, 1) model is selected as appropriate model for prediction of price of tomato in Panipat market.

Hybrid (SARIMA-GARCH) model

The Hybrid model was built in a sequential fashion, first SARIMA model was applied to the original time series and then its residuals was analyzed using GARCH model.

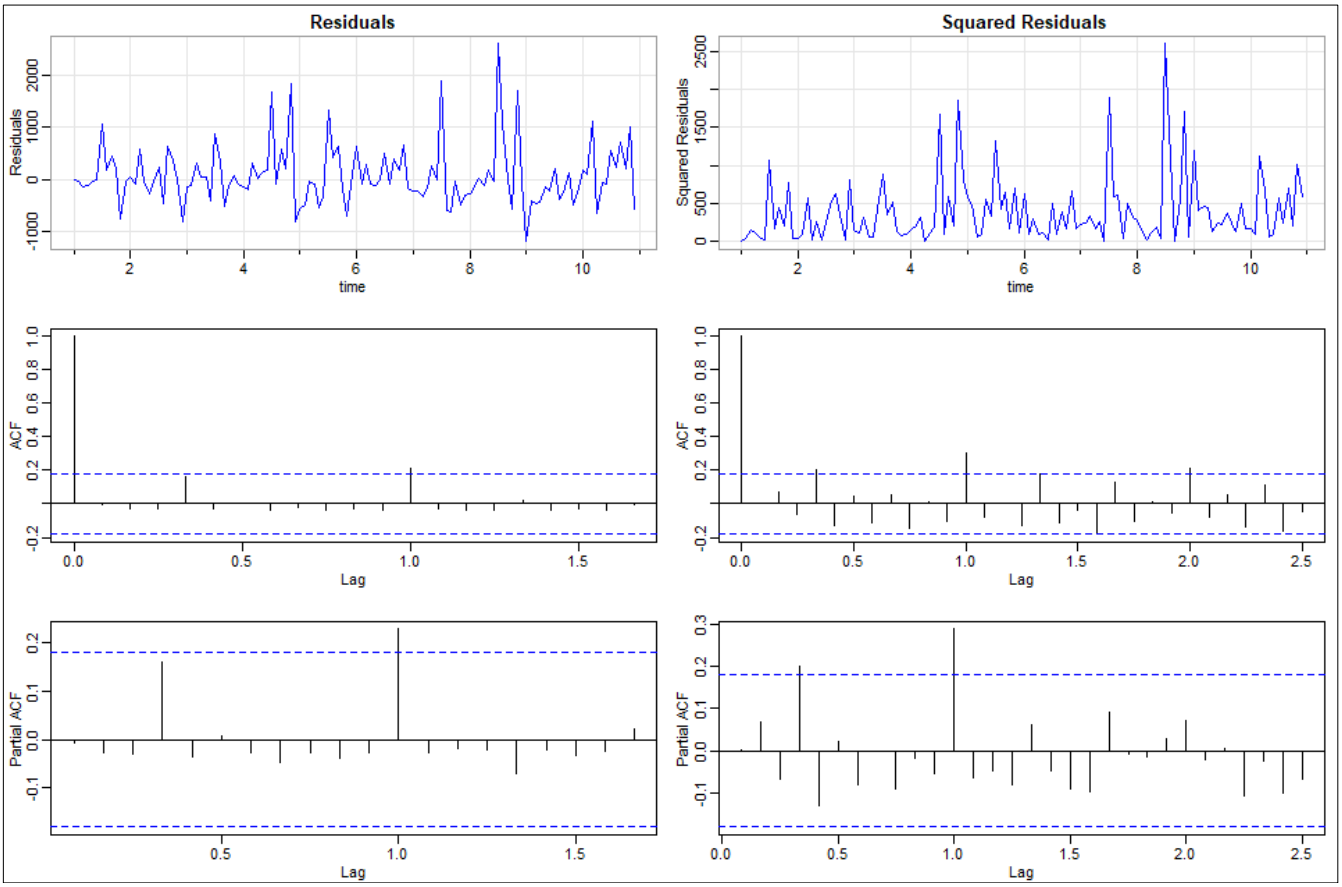


Fig 3: Residuals and squared residuals plots

Figure 3 shows the plots of residuals and squared residuals obtained from best fitted SARIMA model on prices of tomato in Panipat market. It indicates that residuals are not auto-correlated but squared residuals are auto-correlated that means volatility clustering is visible. ACF and PACF plots of residuals reveal that ACF and PACF are not significant at any lags. Also, ACF coefficients are significant at lags 4, 12 and 24, while PACF coefficients are significant at lags 4 and 12 for squared residuals. Thus, the residuals are not auto-correlated but squared residuals are auto-correlated.

Table 4: Estimated parameters of SARIMA (0, 1, 0), (2, 1, 0)$_{12}$-GARCH (1, 1) model

Parameter	Estimate	S. E.	t-value	p-value
SAR1	0.37	0.08	4.19	<0.01
SAR2	0.24	0.09	2.69	<0.01
α_0	224.10	454.23	0.89	0.23
α_1	0.12	0.06	3.01	<0.01
β_1	0.82	0.02	32.93	<0.01
Model selection criteria	AIC	15.56	BIC	15.68
Diagnostic checking	Ljung-Box		ARCH LM	
	Statistic	p-value	Statistic	p-value
	14.98	0.45	10.68	0.55

Table 4, the estimated parameters of selected model for prices (SAR1, SAR2, α_1 and β_1) and arrivals (AR1, AR2, α_1 and β_1) of tomato in Panipat market are given and are found significant at 5% level of significance. Coefficients of ARCH and GARCH (α_1 and β_1) are greater than zero and sum of these coefficients is less than one satisfied the sufficient condition of conditional variance for GARCH model.

Table 5: Forecasting performance of SARIMA (1,1,2)(0,1,1)$_{12}$, ARMA (1,1)-GARCH (0,1) and SARIMA (1,1,2)(0,1,1)$_{12}$ - GARCH (1,1) models

		SARIMA		GARCH		Hybrid	
Month	Observed	Predicted	RD	Predicted	RD	Predicted	RD
Jan-20	2274.09	2196.61	3.41	2056.61	9.56	1996.69	12.20
Feb-20	1334.74	1425.37	-6.79	1158.60	13.20	1245.11	6.72
March-20	1422.76	1506.02	-5.85	1161.87	18.34	1591.33	-11.85
April-20	1422.76	1525.10	-7.19	1289.76	9.35	1421.20	0.11
May-20	691.46	517.79	25.12	898.72	-29.97	807.59	-16.80
June-20	668.42	555.20	16.94	716.02	-7.12	701.62	-4.97
July-20	2006.78	2080.48	-3.67	2242.07	-11.72	2119.80	-5.63
Aug-20	2001.09	2310.33	-15.45	2267.51	-13.31	2231.14	-11.50
Sept-20	2712.35	2799.12	-3.20	2302.35	15.12	2730.34	-0.66

Oct-20	2740.49	2597.08	5.23	2306.72	15.83	2522.44	7.96
Nov-20	2658.78	2752.43	-3.52	2309.38	13.14	2681.09	-0.84
Dec-20	1598.77	2158.31	-35.00	2325.54	-45.46	1785.12	-11.66
MAPE		10.95		16.84		7.57	
SEP		11.66		18.64		8.46	

The MAPE and SEP values for SARIMA, GARCH and Hybrid models shown in Table 5. The hybrid model provides better forecast accuracy in terms of smallest values for MAPE (7.57) and SEP (8.46) as compared to SARIMA (MAPE = 10.95 & SEP = 11.66) and GARCH (MAPE = 16.84 &SEP = 18.64) models

Conclusions

- Time series forecasting is vital for predicting future values based on past data.
- Autoregressive (AR) and Moving Average (MA) models were first formulated by Slutsky (1937) [11] and Yaglom (1955) [12], with Box and Jenkins (1970) [2] integrating these into the ARIMA framework.
- SARIMA models are widely used for modeling seasonality in time series data.
- GARCH models are effective for modeling volatility in time series data.
- Hybrid models, combining SARIMA and GARCH, are valuable for capturing both linear and nonlinear patterns in data.
- Model selection and evaluation are crucial, with criteria like AIC and BIC used for selection and metrics like MAPE and RMSE used for evaluation.
- The selected SARIMA (1, 1, 2), (0, 1, 1)12 model was found appropriate for predicting tomato prices in Panipat market.
- The ARMA (1, 1) - GARCH (0, 1) model was selected for predicting tomato price returns in Panipat market.
- The hybrid SARIMA-GARCH model effectively combines linear and nonlinear components for forecasting.

References

1. Bollerslev T. Generalized Autoregressive Conditional Heteroscedasticity. Journal of Econometrics. 1986;31:307-327.
2. Box GEP, Jenkins GM. Time Series Analysis: Forecasting and Control. San Francisco: Holden-Day; 1970.
3. Chandran KP, Pandey NK. Potato Price Forecasting using Seasonal ARIMA Approach. Potato Journal. 2007;34(1):1-2.
4. Engle RF. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. Econometrica. 1982;50:987-1007.
5. Hakan A, Murat M. An Analysis of Tomato Prices at Wholesale Level in Turkey: An Application of SARIMA Model. Custos E-Agronegocio. 2012;8(4):52-75.
6. Jarque CM, Bera AK. Efficient tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals. Economics Letters. 1980;6(3):255-259.
7. Ljung GM, Box GEP. On a Measure of Lack of Fit in Time Series Models. Biometrika. 1978;65:297-303.
8. Malik V. ARIMA/ GARCH (1,1) Modelling and Forecasting for a GE Stock Price using R. ELK Asia Pacific Journal of Marketing and Retail Management. 2015;8(1):2317-2349.
9. Mallikarjuna HB, Paul A, Noel AS, Sudheendra M. Forecasting of Black Pepper Price in Karnataka State: An Application of ARIMA and ARCH Models. International Journal of Current Microbiology and Applied Sciences. 2019;8(1):1486-1496.
10. Shetty DK, Sumithra, Ismail B. Hybrid SARIMA-GARCH Model for Forecasting Indian Gold Price. International Journal of Multidisciplinary. 2018;3(8):263-269.
11. Slutsky E. The Sommation of Random Causes as the Source of Cycle Processes. Econometrica. 1937;5:105-146.
12. Yaglom AM. Correlation Theory of Processes with Stationary Increments of Order n. American Mathematical Society Translations. 1955;2(8):37-141.
13. Yollanda M, Devianto D. Hybrid Model of Seasonal ARIMA-ANN to Forecast Tourist Arrivals through Minangkabau International Airport. Department of Mathematics, Andalas University, Padang. 2020;3(5):755-765.