Journal of Applied Statistics & Machine Learning
Volume 2; Number 1; 2023 : pp. 79-96

TIME SERIES FORECASTING FOR MARKET PRICES AND ARRIVAL QUANTITIES OF POTATOES IN LUCKNOW

Samapriya Trivedi[1], Shalini Jaiswal[2], Ashok Kumar[3] and Shambhavi Mishra[4*]

[1&2]Research Scholar, Department of Statistics, University of Lucknow, Lucknow, Uttar Pradesh, India
[3]Associate Professor, Department of Statistics, Central University of South Bihar, Gaya, Bihar, India
[4]Assistant Professor, Department of Statistics, University of Lucknow, Lucknow, Uttar Pradesh, India
[*]Corresponding author: shambhavimishra.lko@gmail.com

ARTICLEINFO

Received: 10 February 2022
Revised: 01 March 2023
Accepted: 06 March 2023
Online: 10 March 2023

To cite this paper:
Samapriya Trivedi, Shalini Jaiswal, Ashok Kumar & Shambhavi Mishra (2023). Time Series Forecasting for Market Prices and Arrival Quantities of Potatoes in Lucknow. *Journal of Applied Statistics & Machine Learning*. 2(1): pp. 79-96.

ABSTRACT

In terms of potato production, India ranks in second position worldwide, and the state of Uttar Pradesh produces the majority of them. The wholesale cost and availability of potatoes are crucial factors for farmers, merchants, and customers. The present research attempts to examine the market prices and arriving quantities of potatoes in the wholesale market of Lucknow district, the capital of Uttar Pradesh. The time-series data from January 2011 to December 2022 was collected from the Agricultural Marketing Information Network (AGMARKNET) website, which was started by the Union Ministry of Agriculture, Govt. of India. The compound annual growth rates and correlation coefficient for market prices and arrival quantity are computed and for their better forecasts, four types of time series forecasting models are applied. The Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) models, which are appropriate for stationary time series having clear patterns of seasonality and trend as well as Prophet and Simple Recurrent Neural Network (RNN) models, which are usually used when time series is non-stationary or with complex nonlinear relationships. These models are applied to the univariate time series data, and the predictive accuracies are compared based on Root Mean Square Error (RMSE). The Prophet model came out to be the best-fitted among the four applied models; hence, it is used to forecast the market prices and arrival quantities of potatoes in Lucknow for the next two years, *i.e.*, from January 2023 to December 2024.

Keywords: ARIMA, SARIMA, Prophet, RNN, AIC, RMSE.

1. INTRODUCTION

The Indian economy heavily relies on the agriculture sector and has had substantial growth over the last few years. The agriculture sector constitutes 45.6 percent workforce and accounted for 18.29 percent of the Gross Value Added (GVA) in 2019-20 (Chand & Singh, 2022). In India, where they account for a sizeable portion of the agricultural market, potatoes are a key crop. The Ministry of Agriculture and Farmers Welfare (MoA & FW) estimates that India produced 52.48 million metric tonnes (MMT) of potatoes overall in the years 2020–21. After China, India stands in the second position in terms of potato production in the world. In 2020-21, India produced around 52.48 MMT of potatoes, accounting for approximately 22 percent of the total vegetable production in the country. According to the MoA & FW, the total area for potato cultivation during the year 2020-21 was around 2.13 million hectares. The largest area under potato cultivation was in the state of Uttar Pradesh, which accounted for around 47 percent of the total potato acreage in India, followed by West Bengal, Bihar, and Gujarat.

Potatoes are also one of the most consumed vegetables in India and are an essential part of the Indian diet. It contributes significantly to food security in India and is one of the most consumed vegetables in the country, and its availability and affordability make it a popular food option for many people. It is also a good source of carbohydrates, dietary fibre, vitamins, and minerals. According to the MoA & FW, the per capita consumption of potatoes in India is around 25 kilograms per year, meaning that India consumes around 33 MMT of potatoes per year. They are particularly rich in vitamin C, which is essential for a healthy immune system. The use of potatoes in India is predicted to keep increasing in the coming years due to the expanding population, urbanization, and dietary changes. It's important to mention that potato cultivation in India has also grown recently, reflecting the surging demand for this crop. Furthermore, the government has implemented several initiatives to support potato farming and boost output in response to the rising demand. These initiatives involve subsidizing the purchase of inputs such as seeds, fertilizers, and pesticides and encouraging the adoption of contemporary agricultural technologies and practices.

In the wholesale market, there are two most important factors, one is the arrival quantity of the commodity and another is its price. By closely monitoring the accurate prices and quantities of goods in the wholesale market, the government can take pre-emptive steps and make informed decisions on different policy measures. These may include modifying the

Minimum Support Price (MSP) to guarantee that farmers get a reasonable amount for their produce, as well as setting a Minimum Export Price (MEP) to restrict export prices and compel exporters to sell domestically in order to decrease crop prices. The farmers are also better able to switch between local marketplaces to sell their crops and obtain fair prices for their commodities when pricing information is accessible. The information can be used by the farmers to make decisions regarding the time of marketing. Although, the seasonality factor is the largest barrier to establishing precise projections of prices as well as the arrival of crops in the market. Many approaches have been presented to capture the behavior of crop price and arrival quantity because of the intricacy of the time series data, however, academics have not agreed on the most effective model for vegetable pricing (Xiong et al., 2018). Many linear and nonlinear methods have been established within the time series framework, including the ARIMA, SARIMA, and Generalized Autoregressive Conditional Heteroscedastic (GARCH) models. Numerous studies were done in the past with the goal of forecasting agricultural commodities. For predicting onion rates in Mumbai shops, the SARIMA model is said to perform better than other price forecasting models (Sankaran, 2014), using of SARIMA model for predicting India's meat exports (Paul et al., 2013). In order to forecast wholesale prices in Odisha, several machine-learning techniques were employed, including Support Vector Regression (SVR), Random Forest, *etc*. (Paul et al., 2022). The ARIMA modeling technique to forecast the paddy prices for 5 major states, Uttar Pradesh, Tamil Nadu, Punjab, Delhi, and West Bengal (Kathayat & Dixit, 2021). Forecasting of sale prices on the Italian food wholesaler, in which ARIMA and Long Short-Term Memory (LSTM) models performed better than Convolutional Neural Networks (CNN) and Prophet models (Menculini et al., 2021). The price fluctuations of agricultural goods in Japan were analyzed using recurrent neural network techniques, whose findings indicate that the evaluation criteria are the basis of the selection of a forecasting method, moreover, utilizing a combination of forecasting methods can be more effective (Kurumatani, 2020).

In this paper, the four time-series forecasting models, ARIMA, SARIMA, Prophet, and Simple Recurrent Neural Network (RNN) models are fitted and compared for their predictive accuracies based on accuracy metrics, namely RMSE, for the price and quantity of potatoes in the wholesale market of Lucknow, Uttar Pradesh. The paper is divided into four subsequent parts, Section 2 details the data source, from which the time-series data for the study was obtained. Section 3 elaborates on four forecasting models used in this study. The tables and charts of the results, with their interpretations, are presented in Section 4. The final Section 5 concludes the study, in which the final and best-fitted model, along with some suggestions for extending this research are presented.

2. DATA SOURCE

The website AGMARKNET (*https://agmarknet.gov.in/*) has been used to gather monthly information on the market price (in Rs. /Quintal) and arrival quantity (in tonnes) of potatoes for the Lucknow district of Uttar Pradesh, India, from 1 January 2011 to 31 December 2022. The Union Ministry of Agriculture launched the AGMARKNET in March 2000 under its centrally sponsored Marketing Research as well as Information Network Scheme. The Directorate of Marketing and Inspection (DMI) implements it in collaboration with the State Agricultural Marketing Boards and Agricultural Produce Market Committees (APMCs), and is managed by the National Informatics Centre (NIC).

For the study of potato crop data in Lucknow, the monthly average of market price and arrival quantities from January 2011 to December 2020 was used as the training dataset. Meanwhile, the test period was set to cover January 2021 to December 2022. Several forecasting techniques, including ARIMA, SARIMA, Prophet, and Simple Recurrent Neural Network models, were employed for analysis using the Python v3.10 programming language using Google Colab Notebooks.

3. METHODS

3.1. Autoregressive Integrated Moving Average (ARIMA) Model

A time series is a collection of data points gathered over time, and time series forecasting entails projecting future values for it. Finding patterns in previous data and using them to predict future values accurately is the basic objective of time series forecasting. There are numerous methods available for this purpose, ranging from basic techniques such as moving averages and exponential smoothing to more sophisticated models like ARIMA and neural networks.

The widely used ARIMA model can simulate and predict data from stationary and non-stationary time series, and it is useful in time series forecasting. Its extensive use of the well-known Box-Jenkins technique (Box & Jenkins, 1976) in model construction and numerous statistical features is the cause of its enormous popularity. The three parameters of the ARIMA system are ‘p’, which denotes the number of lagged values of the time series used as predictors, ‘d’, which denotes the number of differenced time series required for the time series to become stationary, and ‘q’, which denotes the number of lagged forecast errors included as predictors. The mathematical equation for '*ARIMA(p,d,q)*' can be written as:

$$y_{(t)} = c + \varphi_1 y_{(t-1)} + \varphi_2 y_{(t-2)} + \ldots + \varphi_p y_{(t-p)} + \theta_1 \varepsilon_{(t-1)} + \theta_2 \varepsilon_{(t-2)} + \ldots + \theta_q \varepsilon_{(t-q)} + \varepsilon_{(t)} \quad (1)$$

where,

$y(t)$: time series value at time 't',

c: constant term (intercept),

$\varphi_1, \varphi_2, \ldots \varphi_p$: auto-regressive coefficients which are past value effect of time series on its current one,

$\theta_1, \theta_2, \ldots . \theta_p$: moving average coefficients, representing past error effect of time series on the current one,

$\varepsilon_{(t)}$: error (residual) term at time 't'.

A series whose statistical characteristics are independent of the time point at which they are observed is referred to as a stationary time series. A popular test utilized to evaluate if the time series is stationary is the Augmented Dickey-Fuller (ADF) test. It tests the null hypothesis that the given time-series lacks a unit root (Fuller & Dickey, 1979). The Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) plots are used to determine the order of auto-regression (p) and moving average (q), respectively. Iterative least squares or maximum likelihood techniques are used to estimate the parameters at this stage.

3.2. Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

The SARIMA model is useful when time series exhibit significant seasonal impacts, as opposed to when there is no strong seasonal trend in the data (Vishwakarma et al., 2020). This model is expressed using six parameters: p, d, q, P, D, and Q. The first three parameters have the same definitions as in the ARIMA model. The parameter 'P' specifies the number of lagged values of the seasonal component that should be included as predictors, and the seasonal component must be differentiated 'D' times for it to become stationary. The parameter 'Q' specifies the number of lagged forecast errors that should be included as predictors for the seasonal component. The mathematical equation for '*SARIMA(p,d,q)(P,D,Q*' can be written as (Goodarzi, 2020):

$$\begin{aligned} y_{(t)} = c + \Phi_1 y_{(t-s)} + \Phi_2 y_{(t-2s)} + \ldots + \Phi_P y_{(t-Ps)} + \varphi_1 y_{(t-1)} + \varphi_2 y_{(t-2)} + \ldots + \varphi_p y_{(t-p)} \\ + \theta_1 \varepsilon_{(t-1)} + \theta_2 \varepsilon_{(t-2)} + \ldots + \theta_q \varepsilon_{(t-q)} + \varepsilon_{(t)} \end{aligned} \quad (2)$$

where:

$y_{(t)}$ time series value at time *'t'*,

c : constant term (intercept),

s : the seasonal period,

$\varphi_1, \varphi_2, \ldots \ldots, \varphi_p$: auto-regressive coefficients, which express the impact of past values of the time series on its present value,

$\Phi_1, \Phi_2, \ldots \ldots, \Phi_P$ seasonal autoregressive coefficients, which indicate the past value effect of the seasonal component on the current one,

$\theta_1, \theta_2, \ldots \ldots . \theta_q$: moving average coefficients, which indicate the past error effect of the time series on the current one,

$\varepsilon_{(t)}$ denotes the error (residual) term at time *'t'*.

3.3. Prophet Model

Facebook Inc. released the Prophet model for time series forecasting in 2017 (Taylor & Letham, 2018). Prophet is an additive regression model that can fit seasonality at the yearly, monthly, and daily levels as well as the effects of holidays in non-linear trends. It also includes several additional features, including the ability to model uncertainty in the trend and seasonal components, the ability to include additional regressors, and handle missing data. It uses a to estimate the parameters of its time series model. It is a non-parametric model, and instead of obtaining a single "best" estimate of the model parameters, it employs a Bayesian regression approach to estimate the posterior distribution of the parameters for the data. Here, the Stan implementation of HMC (Hamiltonian Monte Carlo) is used to sample from the posterior distribution of system parameters. A generalized additive model may be formed for the prophet model, *i.e.*, trend, holidays, and seasonality (Niu et al., 2020):

$$y_{(t)} = g_{(t)} + s_{(t)} + h_{(t)} + \varepsilon_{(t)} \tag{3}$$

where,

$y_{(t)}$: time series value at time 't',

$g_{(t)}$: trend function, which models non-periodic changes in the time series,

$s_{(t)}$: seasonal function, which models periodic changes in the time series,

$h_{(t)}$: holiday function, which models the effects of holidays or other special events on the time series,

$\varepsilon_{(t)}$: error term, which represents the random variation in the time series that is not accounted for by the trend, seasonal, and holiday functions.

The use of a Bayesian approach in Prophet has several advantages. First, it allows for the inclusion of uncertainty in the parameter estimates, which can be propagated through to the forecasts. This means that Prophet can provide not only point forecasts but also uncertainty intervals that reflect the uncertainty in the model parameters. Second, the use of a prior distribution allows for the incorporation of domain knowledge or expert opinion into the model. Finally, the use of MCMC methods allows for flexible modeling of complex and non-linear relationships between the time series and the covariates.

3.4. Simple Recurrent Neural Network (RNN)

Simple RNN, which is a type of neural network architecture is used to analyse sequential data, such as time series or spoken language. Given that it just has one layer of recurrent nodes, it is called "simple" because to its simplicity. It can manage complex nonlinear interactions between past and present values, predicting the current value in a time-series data set using past values, and can identify both short- and long-term dependencies in the time series. Simple RNNs are distinct from regular neural networks in that they can retain information in the form of "memory" from previous steps in the sequence, allowing them to consider the order of the data they're processing. Basic formula of RNN is:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}; \theta) \tag{4}$$

where,

$x^{(t)}$: vector with the time step index 't',

t: time step index, and ranges from 1 to n,

$h^{(t)}$: hidden state vector,

θ : parameters of the function f.

RNNs use feedback loops to incorporate previous time steps' results into the current time step, allowing them to create complex temporal patterns in time series datasets. The input layer receives the input, and it is then passed on to a layer of memory cells or recurrent nodes in the hidden layer. Due to links between the hidden layer and itself, each time step's output can be fed back into the network as input for the next time step. Finally, the output layer produces the predicted output for that time step. The accuracy of RNN predictions depends heavily on the activation function, which uses the current input and previous state to generate output. The loss function is used to measures the difference between the predicted and actual output, and the Simple RNN model is trained to minimize this loss over the data.

3.5. Model Evaluation

We used Akaike's Information Criterion (AIC) to choose the most cost-effective approach for ARIMA and SARIMA for prediction. It is provided by Akaike (1974) and serves as an estimate of the relative quality of statistical approaches for a certain time series.

$$AIC = 2n - 2\ln(\hat{L}) \tag{5}$$

where, L is the model's maximum likelihood function value and n is the number of estimated parameters in the model. We selected the final model as the one with the minimum AIC (Jadhav et al., 2017).

The comparison of ARIMA, SARIMA, Prophet and Simple RNN models was done using MAE (Mean Absolute Error), which measures only the magnitude of the errors and doesn't concern itself with their direction. MAPE (Mean Absolute Percentage Error), which increases linearly with an increase in error. In order to determine if a feature is enhancing the model's prediction or not, the ideal assessment metric is the RMSE, which assesses the average size of mistakes and is concerned with deviations from the actual value. RMSE may be used with a variety of features. The lower the values of these indices, the better the model and its predictions (Guo et al., 2020; Shah V., 2020).

4. RESULTS

For the period of January 2011 to December 2022, mean and standard deviation of market price and arrival quantity of potatoes in Lucknow, along with Compound Annual Growth Rate (CAGR) and correlation coefficients are presented in Table 1.

The CAGR for both the market price (5.2 percent) and arrival quantity (7.7 percent) was found to be significant at a 5 percent level of significance. This shows that the market arrival quantities of potatoes have increased at a greater rate than its prices during the study period. In a report, Pandey et al. (2005) showed that there was a decline in CAGR of potato productivity in Uttar Pradesh from 1980 onward but negative from the 2000s onwards. This was due to farmers' increasing tendency to harvest crops to accommodate the third crop in the year. Also, the correlation coefficient ($r = 0.143$) between the average market price and arrival quantities of

Table 1: Yearly Market Price (Rs./Quintal) and Arrival Quantity (tonnes) of potatoes in Lucknow

Years	*Market Price*	*Market Arrival Quantity*
	Mean ± SD	*Mean ± SD*
2011	541.89 ± 147.64	42.39 ± 43.26
2012	918.29 ± 312.98	93.31 ± 89.29
2013	903.74 ± 250.78	369.83 ± 237.41
2014	1322.24 ± 494.72	207.05 ± 168.31
2015	713.04 ± 123.61	253.67 ± 124.81
2016	1034.75 ± 365.27	195.22 ± 93.16
2017	552.43 ± 151.26	146.07 ± 47.64
2018	1132.01 ± 385.84	117.2 ± 54.25
2019	885.49 ± 245.51	106.65 ± 25.61
2020	1957.68 ± 686.2	102.69 ± 46.68
2021	1011.87 ± 196.88	212.42 ± 72.9
2022	1287.06 ± 355.09	575.13 ± 627.28
CAGR	5.2*	7.7*
r	0.143	

CAGR = Compound Annual Growth Rate, r = Pearson Correlation Coefficient

* Statistically significant ($p < 0.05$)

potatoes in the Lucknow market during the study period was found to be insignificant at a 5% level, indicating no significant relationship between the two variables.

The monthly average of market price and arrival quantity from January 2011 to December 2022 are shown in Figure 1 and Figure 2 respectively. It can be seen that although the arrival quantity of potatoes decreased from March to July, it was significantly higher in last three months of the year. This corresponds to the growth seasons for potatoes in India, where mid-June to mid-July, followed by October and November, are said to be the optimum times to cultivate potatoes. The market prices do not correlate significantly with the availability of the potatoes in the market ($r = -0.507$) as only the price in December, January, and February was lower but it increased with the months.

To fit time series models, it is important to first understand the trend and seasonality components of the market price and arrival quantity of potatoes in Lucknow. This was accomplished by using time-series plots to decompose the data into three components: trend, seasonal, and residual, as depicted in Figure 3 and Figure 4. The time series' long-term movement is represented by the trend component, while the seasonal component

Figure 1: Month-wise distribution of Market Price (Rs./Quintal) of potatoes in Lucknow

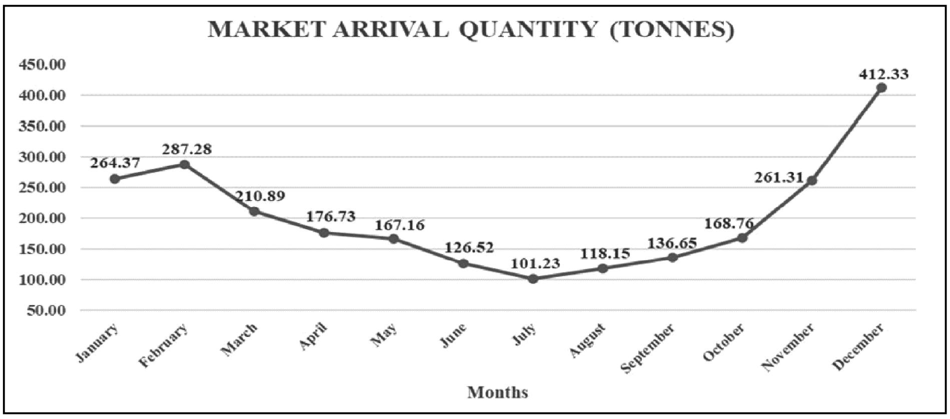

Figure 2: Month-wise distribution of Market Arrival Quantity (tonnes) of potatoes in Lucknow

represents short-term, fixed patterns. The residual component, also known as statistical noise, is the remaining variation after the trend and seasonal components have been accounted for.

The time series in both sets of data exhibits a substantial seasonal component, or variations in a time series that represent intra-year fluctuations that are more or less constant in terms of timing, direction, and size from year to year. For the market price, in Figure 3, it is can be inferred that there is no visible trend present, whereas from Figure 4, it is clear that the time series for the market arrival numbers contains a trend component. Consequently, we must de-trend the time series for the market arrival quantity before applying ARIMA or SARIMA models.

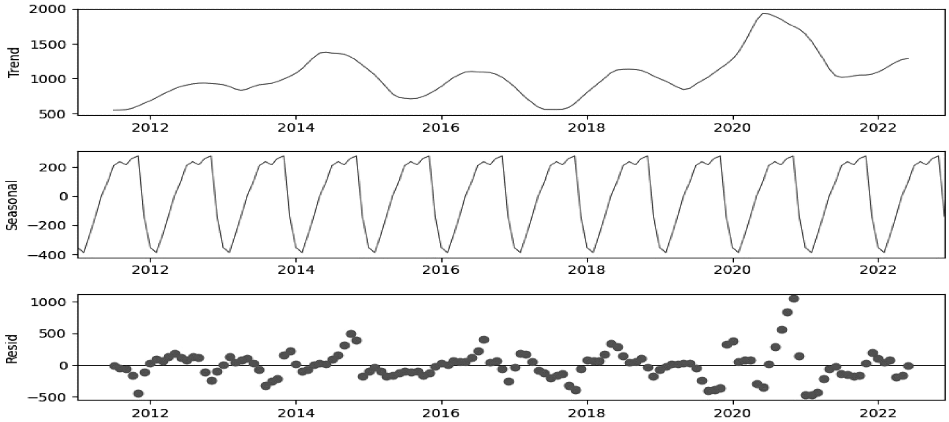

Figure 3: Decomposition of Market Prices (Rs./Quintal) for potatoes in Lucknow

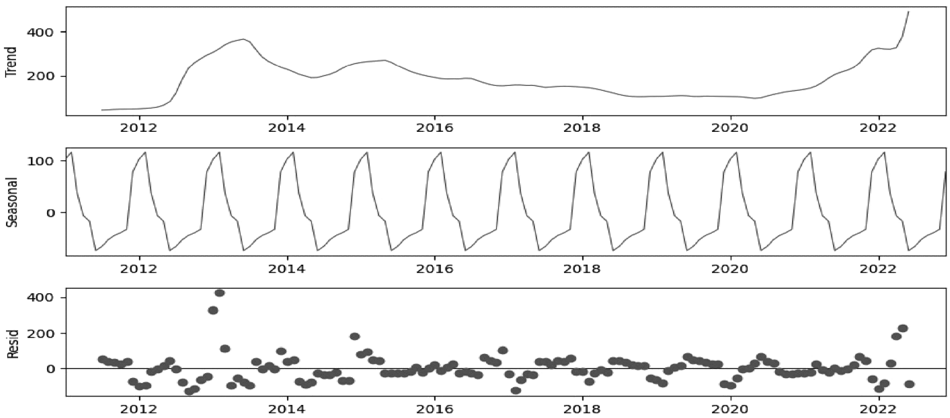

Figure 4: Decomposition of Market Arrival Quantity (tonnes) for potatoes in Lucknow

The stationarity of the time series for the market price and arrival quantities are both examined using the ADF test. The test shows that the market prices time series is stationary (Dickey-Fuller = -4.608, p.= <0.001), and hence no further differencing is required. However, for the market arrival quantity, the test result indicates that the time series is non-stationary (Dickey-Fuller = 0.066, p.=0.964). To make it stationary, first-order differencing (d = 1) is applied, which results in a stationary time series (Dickey-Fuller = -5.976, p.= <0.001).

The next stage in creating an ARIMA and SARIMA model is to choose appropriate values for the AR term, p, and the MA term, q, in the model. This is done after obtaining the differencing value (d) from the ADF test.

For Market Price **For Market Arrival Quantity**

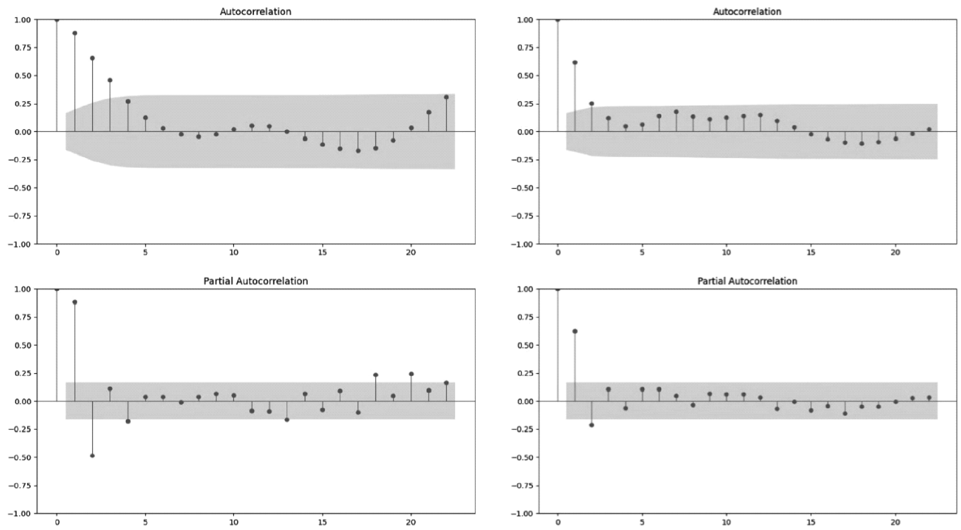

Figure 5: ACF and PACF plots for Market Prices (Rs./Quintal) and Arrival Quantity (tonnes) of potatoes in Lucknow

The auto correlation function (ACF) and partial auto correlation function (PACF) plots of the stationary time series must be analysed in order to ascertain the values of p and q.

The ACF and PACF plots for market price and arrival quantity of potato are shown in Figure 5. For market price, the maximum value of p is 2 and q is 3 whereas for the market arrival quantity, maximum value of p is 2 and q is 2. Now, for the ARIMA and SARIMA time-series forecasting models, different models were applied based on all the possible combinations of values of the parameters. The model having the lowest AIC value has been selected among the different combinations of p, d, q, and P, D, Q values.

Table 3: ARIMA and SARIMA Models for Market Price (Rs./Quintal) and Arrival Quantity (tonnes) of potatoes in Lucknow

Market Price		*Market Arrival Quantity*	
Model	*AIC*	*Model*	*AIC*
ARIMA (2, 0, 0)	1641.06	ARIMA (2, 1, 1)	1380.89
ARIMA (2, 0, 1)	1642.29	ARIMA (2, 1, 0)	1391.88
ARIMA (1, 0, 3)	1644.57	ARIMA (0, 1, 2)	1393.81
ARIMA (2, 0, 3)	1644.12	ARIMA (2, 1, 2)	1383.06
SARIMA (1, 0, 3) (2, 0, 2)12	1611.26	SARIMA (1, 1, 2) (1, 1, 1)12	1246.37
SARIMA (2, 0, 1), (2, 0, 2)12	1613.35	SARIMA (0, 1, 2) (1, 1, 1)12	1249.47
SARIMA (2, 0, 3), (1, 0, 0)12	1653.68	SARIMA (2, 1, 0) (2, 1, 2) 12	1253.30
SARIMA (2, 0, 3) (2, 0, 3)12	1617.29	SARIMA (2, 1, 2) (2, 1, 2)12	1252.82

Based on the AIC values given in Table 3, for the market price of potatoes, the best ARIMA model is ARIMA (1, 0, 2) while for the SARIMA model, it is SARIMA (1, 0, 3) (2, 0, 2)12. Whereas, for the market arrival quantity of potatoes, the best-fitted ARIMA model is ARIMA (2, 1, 1) while the best-fitted SARIMA model is SARIMA (1, 1, 2) (1, 1, 1)12.

The Prophet Model has been fine-tuned by turning *off* the weekly and daily seasonality and using a different combination of Changepoint Prior Scale values (0.01, 0.001, 0.05, 0.025, 0.25, 0.50) and Seasonality Prior Scale values (1.0, 2.0, 5.0, 8.0, 10.0), which are parameters to change the flexibility of the automatic change-point selection and the strength of the seasonality model respectively. The changepoint prior scale value of 0.001 and seasonality prior scale value of 2.0 were found to produce the lowest RMSE values for the market price, while the changepoint prior scale value of 0.001 and seasonality prior scale value of 5.0 were found to produce the lowest RMSE values for the market arrival quantity of potatoes in Lucknow. These two models are therefore chosen as the best-fitted Prophet models for the market prices and arrival quantities of potatoes in Lucknow.

The fourth model, Simple RNN is fitted using the '*Keras deep learning library*' for both the market arrival quantities and prices of potatoes. For training the model for both market prices and arrival quantities of potatoes in Lucknow, the sequential model is used, with a single dense layer to produce a single value at the output. The model has 24 nodes (or neurons) and the dropout rate was set at 0.3, *i.e.*, which means that 30 percent of the connections will be randomly dropped out during training to prevent overfitting. The *'adam'* optimizer and *'mse'* loss function, which calculates the difference between predicted and actual values, were used to train the neural networks.

Table 4 provides us with the model evaluation metrics, namely RMSE, MAE, and MAPE. For both the market price and arrival quantity of potatoes,

Table 4: Model Evaluation

	RMSE	*MAE*	*MAPE (%)*
For Market Price			
ARIMA (2, 0, 0)	385.282	315.890	25.583
SARIMA (1, 0, 3) (2, 0, 2)12	449.627	421.361	41.679
Prophet	296.542	242.205	24.009
Simple RNN	425.869	344.489	28.771
For Market Arrival Quantity			
ARIMA (2, 1, 1)	480.558	237.927	41.701
SARIMA (1, 1, 2) (1, 1, 1)12	488.338	261.300	51.072
Prophet	430.118	222.968	52.530
Simple RNN	462.893	219.978	40.099

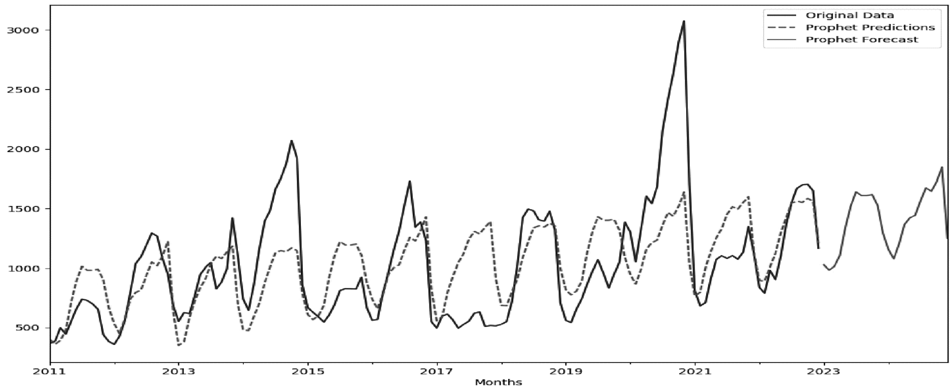

Figure 6: Actual, Predicted and Forecasted values for Market Price (Rs. / Quintal) of potatoes in Lucknow

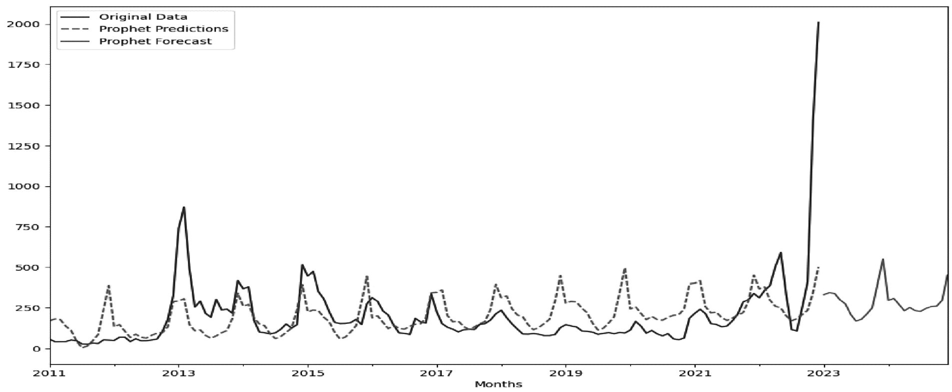

Figure 7: Actual, Predicted and Forecasted values for Market Arrival Quantity (tonnes) of potatoes in Lucknow

the Prophet model has the lowest RSME and MAE values, although the MAPE for market arrival quantity was not the smallest among the four models. Since, RMSE can be considered as the best measure of accuracy for a predictive model, hence, the Prophet models are selected as the best forecasting models for market price and arrival quantity of potatoes in Lucknow.

Using these two forecasting models, the forecast for market prices and arrival quantities of potatoes in Lucknow for the next 24 months (*i.e.*, 2 years) is done with the values shown in Table 5, and the graphs for both depicted in Figure 6 and Figure 7 respectively. In both figures, the blue line represents the original data from the period January 2011 to December 2022 and the green line shows the forecasted values using the Prophet model for the period January 2023 to December 2024.

Table 5: Forecasted values for Market Prices (Rs./Quintal) and Arrival Quantity (tonnes) of potatoes in Lucknow

Time	*Market Price*	*Market Arrival Quantity*
January 2023	1026.76	332.17
February 2023	984.67	343.49
March 2023	1016.59	339.43
April 2023	1104.58	299.62
May 2023	1338.62	272.71
June 2023	1521.20	208.58
July 2023	1640.29	166.94
August 2023	1609.84	177.95
September 2023	1610.07	209.91
October 2023	1617.70	248.24
November 2023	1526.43	396.11
December 2023	1295.27	550.88
January 2024	1157.81	295.69
February 2024	1075.32	308.65
March 2024	1199.06	268.54
April 2024	1365.35	230.11
May 2024	1421.91	250.48
June 2024	1443.17	231.17
July 2024	1567.11	226.74
August 2024	1675.68	244.35
September 2024	1645.05	256.66
October 2024	1727.32	259.80
November 2024	1847.78	295.64
December 2024	1249.13	451.49

The residuals are plotted for both the models in Figure 8, representing histogram and quantile-quantile (q-q) plot. From the histogram, the normal density curve for the market price is close to bells-shaped curve, but for the market arrival quantities, the distribution of residuals is not exactly normally distributed. But, since the q-q plots of both the market prices and arrival quantities of potatoes have the data points primarily concentrated around the reference line and hence, it can be inferred that the standard errors are essentially constant with respect to mean and variance.

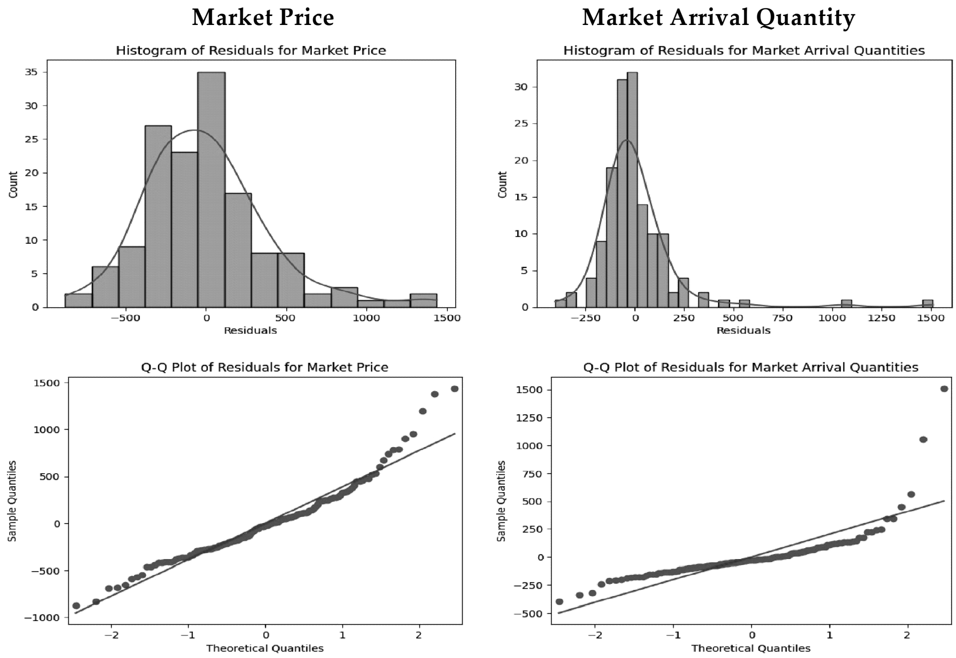

Figure 8: Residual Diagnostic for Market Prices (Rs./Quintal) and Arrival Quantity (tonnes) of potatoes in Lucknow

5. CONCLUSION

In our study, for both the market prices (in Rs./Quintal) and arrival quantity (in tonnes) of potatoes in Lucknow, Uttar Pradesh, the four time-series forecasting models, ARIMA, SARIMA, Prophet and Simple RNN were applied on univariate data for the period of January 2011 to December 2022. The criterion of selecting the best-fitted model was based on several model evaluation techniques, among which RMSE, which is considered as the most appropriate accuracy metric for comparing different types of models as it penalizes large errors. Among the four applied models, the best-fitted model was the Prophet model, and hence, the forecast for the next 24 months (*i.e.*, 2 years), from January 2023 to December 2024 has been done using it for both the market price and arrival quantity of potatoes in Lucknow. The best fitting of the Prophet model for both commodities shows that the Bayesian regression approach used in the model, which takes into account non-linear trends with seasonality, can be used to efficiently predict these two commodities of potatoes, as well as other commodities of the wholesale crop market. Also, the Simple RNN model, which accounts for non-linear trends in time series, trailed the Prophet model. This can be attributed to the Simple RNN model's "vanishing gradient problem," which occurs when the gradient of the error function with respect to the network parameters becomes very small, making it challenging to update the weights during training. This can be addressed by using more advanced types of RNNs, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), which could result in a better fitting of the time-series data available for studying commodities like potato, wheat, rice *etc*. For a deeper analysis, data from multiple markets in Uttar Pradesh can be taken into account for a longer period of time, as well as different factors impacting the production and sale of potatoes, such as temperature, rainfall, cultivated area, *etc*. can be studied for the effect on prices and quantities.

References

Akaike, H. (1974). A new look at the statistical model identification. *Springer Series in Statistics*, 215–222. https://doi.org/10.1007/978-1-4612-1694-0_16

Box, G.E.P. & Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, Revised Edition, San Francisco: *Holden Day*. ISBN 0816211043.

Chand, R., & Singh, J. (2022). Workforce Changes and Employment Some Findings from PLFS Data Series. Retrieved from https://www.niti.gov.in/sites/default/files/2022-04/Discussion_Paper_on_Workforce_05042022.pdf.

Dickey, D., & Fuller, W. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, *74*(366), 427-431. https://doi.org/10.2307/2286348

Goodarzi, M. (2020). Application and performance evaluation of time series, neural networks and Hartt models in predicting groundwater level changes, Najafabad Plain, Iran. *Sustainable Water Resources Management*, *6*, 67. https://doi.org/10.1007/s40899-020-00427-2

Guo, Y., Feng, Y., Qu, F., Zhang, L., Yan, B., & Lv, J. (2020). Prediction of hepatitis E using machine learning models. *PLOS ONE*, *15*(9), e0237750. https://doi.org/10.1371/journal.pone.0237750

Jadhav, V., Chinnappa B.V.R., & Gaddi, G.M. (2017). Application of ARIMA model for forecasting agricultural prices. *Journal of Agricultural Science and Technology*. 19(5). 981-992.

Kathayat, B., & Dixit, A. (2021). Paddy market price forecasting in India using ARIMA mode. *Journal Of Crop and Weed*, *17*(1), 48-55. https://doi.org/10.22271/09746315.2021.v17.i1.1405

Kurumatani, K. (2020). Time series forecasting of agricultural product prices based on recurrent neural networks and its evaluation method. *SN Applied Sciences*, 2(8). 1434. https://doi.org/10.1007/s42452-020-03225-9

Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). Comparing prophet and deep learning to Arima in forecasting wholesale food prices. *Forecasting*, 3(3), 644–662. https://doi.org/10.3390/forecast3030040

Niu, H., Omitaomu, O. A., & Cao, Q. C. (2020). Machine Committee framework for Power Grid Disturbances analysis using synchrophasors data. *Smart Cities*, *4*(1), 1–16. https://doi.org/10.3390/smartcities4010001

Pandey, N.K., Rana, R.K., Pandit, A. and Chandran, K.P. (2005). Contribution of Uttar Pradesh towards Indian potato, *Potato J*. 32(3-4): 121-122

Paul, R., Panwar, S., Sarkar, S., Kumar, A., Singh, K., Farooqi, S., & Choudhary, V. (2013). Modelling and Forecasting of Meat Exports from India. *Agricultural Economics Research Review*, *26*(2), 249 - 255. https://doi.org/10.22004/ag.econ.162149

Paul, R., Yeasin, M., Kumar, P., Kumar, P., Balasubramanian, M., & Roy, H. et al. (2022). Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. *PLOS ONE*, *17*(7), e0270553. https://doi.org/10.1371/journal.pone.0270553

Sankaran, S. (2014). Demand forecasting of fresh vegetable product by seasonal ARIMA model. *International Journal of Operational Research*, *20*(3), 315-330. https://doi.org/10.1504/ijor.2014.062453

Shah V. (2020). A Comparative Study of Univariate Time-series Methods for Sales Forecasting. UWSpace. http://hdl.handle.net/10012/15488

Taylor, S., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, *72*(1), 37-45. https://doi.org/10.1080/00031305.2017.1380080

Vishwakarma, A., Singh, A., Mahadik, A., & Pradhan, R. (2020). Stock Market price Prediction Using SARIMA and Prophet Machine Learning Model. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, *9*(1). Retrieved from http://ijarsct.co.in/Paper315.pdf

Xiong, T., Li, C., & Bao, Y. (2018). Seasonal forecasting of agricultural commodity market price using a hybrid STL and ELM method: Evidence from the vegetable market in China. *Neurocomputing*, *275*, 2831-2844. https://doi.org/10.1016/j.neucom.2017.11.053