International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217440

Vegetable Price Prediction using ARIMA

Sindhuja T[1]*, Sakhi Chawda[1], Parimala Kanaga Devan Kailasam[2]

[1] UG Scholar, Computer Science and Engineering, Easwari Engineering College, Anna University, Chennai, India

[2] Associate Professor, Computer Science and Engineering, Easwari Engineering College, Anna University, Chennai, India

*Corresponding Author: sindhuja.togarrathi@gmail.com

Article Info
Volume 7, Issue 4
Page Number: 176-183

Publication Issue :
July-August-2021

Article History
Accepted : 06 July 2021
Published : 13 July 2021

ABSTRACT

Agriculture is the major occupation of India. The farmers who are the backbone of the country are suffering in utter poverty. This is because they are unaware of the facts that happen in the market. Thereby, they sell their crops at a price much lower than the actual cost. Analyzing data over a time period regularly will lead to various insights and conclusions. These insights can pave way to understand the prices better. Hence, this system suggests ARIMA approach to develop a forecast model and predict, by considering the seasonality in prices over a period of time.

Keywords : time-series, prediction, forecasting, ARIMA

I. INTRODUCTION

Today, India ranks second worldwide in farm output. Agriculture industry contributes about 50% to the economic growth of the country. It is a major contribution to all other industries such as tectiles where the raw materials such as cotton is obtained from plants. This contribution makes it one of the most sought after industries for the raw material. India exported $38 billion worth of agricultural products in 2013, making it the seventh largest agricultural exporter worldwide and the sixth largest net exporter. Most of its agriculture exports serve developing and least developed nations. Indian agricultural/horticultural and processed foods are exported to more than 120 countries, primarily in the Middle East, Southeast Asia, SAARC countries, the EU and the United States.

The organisation of the paper is as follows- literature survey conducted on various forecasting techniques followed by results of ARIMA algorithm, performance of ARIMA, conclusion, future work and references.

II. LITERATURE SURVEY

[1] explains the challenges of forecasting crude oil price due to its high volatility. Since crude oil is the world's leading fuel, and its prices has a global impact on the environment and the economy the prediction method used should be yield high accuracy hence, a new machine learning paradigm called stream learning is used. In stream learning the model is continuously updated as it captures the changing pattern in oil price with small overhead constant. To prove the efficiency of stream learning, the model is

compared to three different models: (1) heuristic approaches; (2) econometric models; and (3) machine learning techniques. Heuristics approach included professional and survey forecasts, which are mainly based on professional knowledge, judgments, opinion and intuition. Econometric model included autoregressive moving average (ARMA) models and vector autoregressive (VAR) models. Machine learning techniques included artificial neural networks (ANN) and support vector machine (SVM). A new stream learning approach was developed to handle applications where continuous data streams are generated from non-stationary processes. Like any machine learning data set, the data is divided into training set and test set. For such an approach it is assumed that both the data sets are homogeneous. Hence, stream learning takes into account recent data history and is updateable.

[2] explains about the most popular measure for accurate forecasting-mean absolute percentage error(MAPE).A significant comparison is done between two forecasting measures, mean absolute percentage error and mean arctangent absolute percentage error(MAAPE) which was developed by looking at mean absolute percentage error from a different perspective. Essentially, MAAPE is a slope as an angle, while MAPE is a slope as a ratio. MAAPE preserves the idea of MAPE but overcomes the problem of division by zero by using bounded influences. MAPE measures the error in terms of percentage, it calculates the number of unsigned percentage errors. Although MAPE is used highly due to its properties of f scale-independency and interpretability it has its issues too, it produces infinite or undefined values for zero or close-to-zero which is a common occurrence in some fields, a very large percentage error is given if the actual values are very small(one or less than one). The MAPE, as a percentage, only makes sense for values where divisions and ratios make sense. It doesn't make sense to calculate percentages of temperatures, for instance, so you shouldn't use the MAPE to calculate the accuracy of a temperature forecast. MAPE allows us to compare forecasts of different series on different scales.

[3] Electricity Price Forecasting (EPF) looks ahead and speculates the direction EPF will or should take in the next decade or so. Five major topics are discussed about EPF (1) Fundamental price drivers and input variables, (2)Beyond point forecasts, (3) Combining forecasts, (4) Multivariate factor models and (5)The need for an EPF-competition. A key point in EPF is the appropriate selection of input variables. On the one hand, the electricity price exhibits seasonality at the daily and weekly levels, and the annual level to some extent. On the other hand the electricity spot price is dependent on a large set of fundamental drivers, including system loads (demand, consumption figures), weather variables (temperatures, wind speed, precipitation, solar radiation), fuel costs (oil and natural gas, and to a lesser extent coal) etc. In Beyond point forecasts we look into interval forecasts which are associated with random variables, density forecasts and threshold forecasting. Combining forecasts combines several forecasts together which shows significantly better results than individual forecasts. There is a need for EPF-competition due to different datasets, different software implementations of the forecasting models and different error measures, but also to the lack of statistical rigor in many studies.

[4] Using data mining techniques for stock prediction to help financial investors to make subjective decisions. One such technique is artificial neural networks (ANN). In ANN the use of technical analysis variables for stock prediction is prevalent. It also explains a hybridized approach which integrates the use of variables of technical and fundamental analysis of the stock market for better prediction and improvement of existing approaches. ANN is gaining heavy attention due to its ability to learn and detect relationships among nonlinear variables. It also does a

deep analysis of large sets of data that usually have a tendency to fluctuate within a short span of time. One of the major drawbacks of ANN is too many hidden nodes and it consumes a lot of time. The focus here is to improve the prediction by using a hybridized method. The technical analysis variables are the core stock market indices which include current stock price, opening price, closing price, volume, highest price and lowest price., while the fundamental analysis variables are company performance indices such as price per annual earning, rumor/news, book value and financial status etc.

[5] explains the working of an Artificial Neural Networks(ANN) with respect to the forecasting of load demand of electricity. This technique is used for short-term load forecasts. The prediction was done on the active hourly variations of power collected from Renigunta substation A.P., India over a period of one month. For training the ANN, active powers were taken as the input quantities and obtained respective active powers for the corresponding day as the output. The algorithm comprises a network with 5 nodes as input, 1 output node and 21 hidden nodes. The process of training involved setting up the nodes' weights with random values between 0.5 to 5.5. Since ANN is a Back-Propagation Network (BPN), for every input vector an output vector is obtained and the error between the desired value and the network output is calculated. This error is used to adjust the weights of the network in a way to minimize the error. The error is back propagated until the error value obtained is considerably low. The issue with ANN is that it requires a fine-tuned architecture to achieve better accuracy.

[6] forecasts short-term prices for electricity using fuzzy neural networks (FNN). This FNN comprises a hypercubic training mechanism with feed-forward architecture. It is designed by combining fuzzy logic with a learning algorithm that models nonstationary behaviour and outliers for the prices presented. The FNN comprises Neural Network (NN) and fuzzification method. A feedforward architecture is used to represent NN that has an input layer, hidden layer and output layer. The input features are propagated to the hidden layers which obtains a vector of weighted inputs. The learning process of FNN involves two tasks by adjusting the parameters: the classification of input vector space based on training samples into hypercubes and, implementing each functional relationship in one class. A comparative study was carried out among DR, TF, MLP, RBF, ARIMA, wavelet-ARIMA and FNN. The performance of FNN was greater than other models. But for FNN, the relationship among various factors has to be assessed in order to determine an accurate forecast, which in this case is not considered.

[7] It is a review on the past 25 years of time series forecasting. It explains over 12 forecasting methods along with their issues and how one overcomes the issue of the other. Starting with Exponential smoothing which shoothens time series data using exponential window function, its drawback is It produces forecasts that lag behind the actual trend. While linear exponential smoothing models are all special cases of ARIMA models, the non-linear exponential smoothing models have no equivalent ARIMA counterparts. There are also many ARIMA models that have no exponential smoothing counterparts. Seasonality, the oldest approach to handling seasonality in time series is to extract it using a seasonal decomposition procedure such as the X-11 method. Seasonality generally cannot be identified until the trend is known, however a good estimate of the trend cannot be made until the series has been seasonally adjusted. Therefore X11 uses an iterative approach to estimate the components of a time series. As a default, it assumes a multiplicative model. State space and structural models and the Kalman filter, used in the early 1980's where statisticians used state space models for time series. State space models provide a unifying framework in which any linear time series model can be written.

As years went on several other models developed such as Nonlinear models, Long memory models, ARCH/GARCH models, Count data forecasting, Combining and Prediction intervals and densities.

[8] Explains the advantages of using exponentially weighted moving average that is used to smoothen random functions that have the following desirable properties: (1) declining weight is put on older data, (2) it is extremely easy to compute, and (3) minimum data is required. This paper utilizes these desirable properties both to smooth current random fluctuations and to revise continuously seasonal and trend adjustments. The flexibility of the method combined with its economy of computation and data requirements make it especially suitable for industrial situations in which a large number of forecasts are needed for sales of individual products. The Moving Average model takes a data set with variation and creates another data set with less variation, or a smoothed data set by aggregating several periods of data. The routines are designed to remove seasonal and random noise variation within time series. Applying this routine repeatedly would result in removal of cyclic variation and left with a combination of trend and some cyclic behaviour. The smoothing effect of the moving average model provides for a "cleaner" data set, which may or may not help in estimating the future level of a variable. The concurrent disadvantage of the greater sensitivity of the EMA is that it is more vulnerable to false signals and getting whipsawed back and forth.

[9] puts forth two accurate and efficient time series approaches: dynamic regression and transfer function models to predict the electricity prices for the next day. The price forecasting in electricity markets tends to maximize the benefits as well as utilities for both producers and consumers. A hypothetical probability model, which is used to represent data, is built with an assumption that the prices are considered after fixed intervals of time. The procedure of analysis is carried out by identification of the model assuming certain hypotheses, followed by estimation of parameters. Then, the model can be used to forecast if the hypotheses are validated, else the model is further refined. The validation of the model is done by checking for randomness of the residuals obtained from autocorrelation plots and partial autocorrelation plots. Both dynamic regression model and transfer function model are used to overcome the serial correlation problem. Initially, the error term is assumed to be from a series with zero mean and constant variance (white noise process). The efficiency of this method is improved based on the selection of parameters which achieves an uncorrelated set of errors. In the final step of validation, the series of residuals is said to be passed if it follows a white noise process. The price and demand series are assumed to be stationary i.e., constant mean and variance. The selection of parameters is done by logarithmic transformation on price and demand series.

[10] Short-term forecasting using Artificial Neural Networks (ANN) comprises three-layered ANN with back-propagation. The first layer being the input layer, hidden layer and one output layer. The training time series data is presented to the network for a certain number of training cycles and its output is compared with the known outputs and an error is generated called Mean Square Error (MSE). MSE is the measure of distance between network output and known target which is back propagated to the input layer. Every cycle is aimed at minimizing the error rate. The learning process is carried out by adjusting the weights that connect input and hidden layers. The epochs are carried out until the error falls below a certain predefined threshold. After all epochs are carried out, the network is ready for prediction and generalisation. The two parameters, learning rate and momentum are used for RMS error minimization. The performance of ANN is reduced if there is any high variation in the data which makes it difficult to evaluate. In the data considered here, there is a high

variation in the prices as the dataset is relatively small.

III. PROPOSED SYSTEM

ARIMA

The ARIMA process can be extended to include seasonal terms, giving a non-stationary seasonal ARIMA (SARIMA) process. Seasonal ARIMA models are powerful tools in the analysis of time series as they are capable of modelling a very wide range of series. Much of the methodology was pioneered by Box and Jenkins in the 1970's.

Series may also be non-stationary because the variance is serially correlated (technically known as conditionally heteroskedastic), which usually results in periods of volatility, where there is a clear change in variance. This is common in financial series, but may also occur in other series such as climate records. One approach to modelling series of this nature is to use an autoregressive model for the variance, i.e. an autoregressive conditional heteroskedastic (ARCH) model.

A seasonal ARIMA model uses differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects. As with lag 1 differencing to remove a trend, the lag s differencing introduces a moving average term. The seasonal ARIMA model includes autoregressive and moving average terms at lag s. The seasonal ARIMA(p, d, q)(P, D, Q)s model can be most succinctly expressed using the backward shift operator

$$\Theta_P (B^s)\ \theta_p(B)(1 - B^s)^D\ (1 - B)^d\ x_t = \Phi_Q(B^s)\ \phi_q(B)w_t \tag{2}$$

where Θ_P , θ_p, Φ_Q, and ϕ_q are polynomials of orders P, p, Q, and q, respectively.

In general, the model is non-stationary, although if D = d = 0 and the roots of the characteristic equation (polynomial terms on the left-hand side of Equation (2)) all exceed unity in absolute value, the resulting model would be stationary. Some examples of seasonal ARIMA models are:

(a) A simple AR model with a seasonal period of 12 units, denoted as

ARIMA(0, 0, 0)(1, 0, 0), is $x_t = \alpha x_{t-12} + w_t$. Such a model would be appropriate for monthly data when only the value in the month of the previous year influences the current monthly value. The model is stationary when $|\alpha^{-1/12}| > 1$.

(b) It is common to find series with stochastic trends that nevertheless have seasonal influences. The model in (a) above could be extended to

$$x_t = x_{t-1} + \alpha x_{t-12} - \alpha x_{t-13} + w_t$$

Rearranging and factorising gives

$(1 - \alpha B^{12})(1 - B)x_t = w_t$ *or* $\Theta_1 (B^{12})(1 - B)x_t = w_t$, which, on comparing with Equation (2), is ARIMA(0, 1, 0)(1, 0, 0). This model could also be written

$\Delta x_t = \alpha \Delta x_{t-12} + w_t$, which emphasises that the change at time t depends on the change at the same time (i.e., week) of the previous year. The model is non-stationary since the polynomial on the left-hand side contains the term (1 − B), which implies that there exists a unit root B = 1.

(c) A simple quarterly seasonal moving average model is

$x_t = (1 - \beta B^4)w_t = w_t - \beta w_{t-4}$. This is stationary and only suitable for data without a trend. If the data also contain a stochastic trend, the model could be extended to include first-order differences, $x_t = x_{t-1} + w_t - w_{t-4}$,

which is an ARIMA(0, 1, 0)(0, 0, 1) process. Alternatively, if the seasonal terms contain a stochastic trend, differencing can be applied at the seasonal period

to give $x_t = x_{t-4} + w_t - \beta w_{t-4}$, which is ARIMA(0, 0, 0)(0, 1, 1).

Differencing at lag s will remove a linear trend, so there is a choice whether or not to include lag 1 differencing. If lag 1 differencing is included, when a linear trend is appropriate, it will introduce moving average terms into a white noise series. As an example, consider a time series of period 4 that is the sum of a linear trend, four additive seasonals, and white noise:

$$x_t = a + bt + s_{[t]} + w_t$$

where [t] is the remainder after division of t by 4, so $s_{[t]} = s_{[t-4]}$. First, consider first-order differencing at lag 4 only. Then,

$$\begin{aligned}(1 - B^4)x_t &= x_t - x_{t-4}\\ &= a + bt - (a + b(t - 4)) + s_{[t]} - s_{[t-4]} + w_t - w_{t-4}\\ &= 4b + w_t - w_{t-4}\end{aligned}$$

Formally, the model can be expressed as ARIMA(0, 0, 0)(0, 1, 1) with a constant term 4b. Now suppose we apply first-order differencing at lag 1 before differencing at lag 4. Then,

$$\begin{aligned}(1 - B^4)(1 - B)x_t &= (1 - B^4)(b + s_{[t]} - s_{[t-1]} + w_t - w_{t-1})\\ &= w_t - w_{t-1} - w_{t-4} + w_{t-5}\end{aligned}$$

which is a ARIMA(0, 1, 1)(0, 1, 1) model with no constant term.

The below two are the sample crops that have been considered and the results of it are presented in terms of ACF, PACF plots, differentiation and finally, the forecasts.

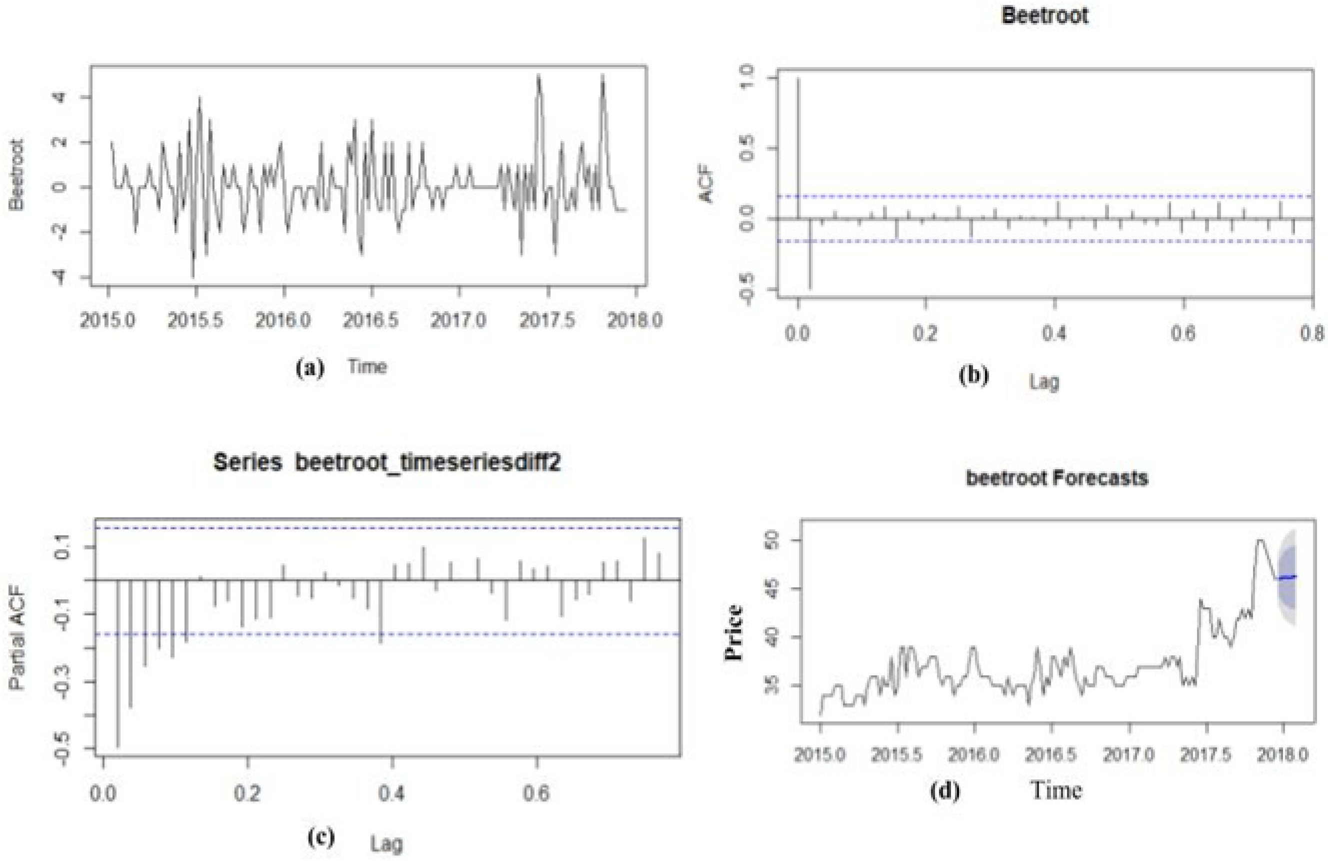

Figure 4.5 ARIMA-Beetroot

(a) Differenced time series-Beetroot (b) ACF Plot-Beetroot (c) PACF Plot- Beetroot (d) ARIMA Forecast-Beetroot

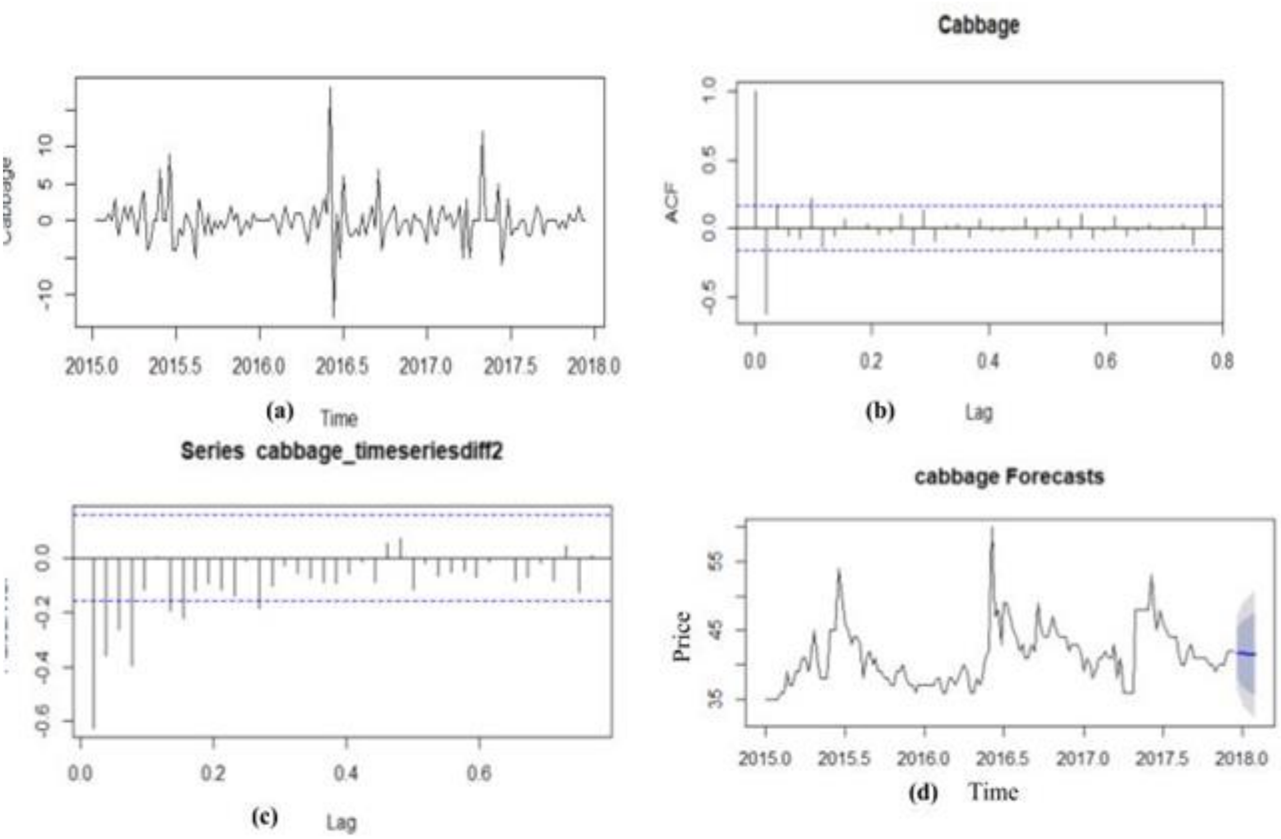

Figure 4.6 ARIMA-Cabbage
(a) Differenced time series-Cabbage (b) ACF Plot-Cabbage (c) PACF Plot- Cabbage
(d) ARIMA Forecast-Cabbage

IV. PERFORMANCE OF ARIMA

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation. It usually expresses accuracy as a percentage, and is defined by the formula:

$$\mathrm{M} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

where A_t is the actual value and F_t is the forecast value.

The difference between At and Ft is divided by the Actual value A_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. Multiplying by 100 makes it a percentage error.

Crop/MAPE	ARIMA(%)
Capsicum	6.19
Beetroot	0.93
Carrot	27.45
Cabbage	6.62
Lettuce	2.95
Onion-Red	4.09
Average	8.04

V. CONCLUSION

Farmers have to face a lot of difficulties and obstacles in their occupation. They are usually kept on the dark side of the consumer market, they are unaware of the rise and fall of the prices of their crop, hence our proposed system not only enlightens them about the prices but also predicts future prices which will help them prepare well for the next harvest. Farmers feed the nation and it is our duty as engineers to use our knowledge and resources to uplift them.

VI. FUTURE WORK

Our vision for this system is that it reaches far and wide throughout the nation, to do so the system can be made multi lingual for easy access. Various external factors that affect the price can also be incorporated for a better understanding of the fluctuation in prices.

VII. REFERENCES

[1]. Shuang Gao, Yalin Lei, A new approach for crude oil price prediction based on stream learning, GeoScience Frontiers, Elsevier, Vol 8, Issue 1, pp:183-187, 2017

[2]. Sungil Kima, Heeyoung Kimb, A New Metric Of Absolute Percentage Error For Intermittent Demand Forecasts, International Journal of Forecasting, Vol 32, Issue 3, Elsevier, pp:669-679, 2016

[3]. Rafał Weron, A look into the future of ‘electricity price forecasting’, International Journal of Forecasting, Vol 30, Issue 4, Elsevier, pp:1030-1081, 2014,

[4]. Adebiyi Ayodele A., Ayo Charles K., Adebiyi Marion O., and Otokiti Sunday O., Stock Price Prediction using Neural Network with Hybridized Market Indicators, Journal of Emerging Trends in Computing and Information Sciences, Vol 3, Issue 1, pp:1-9, 2012

[5]. T. Gowri Manohar and V. C. Veera Reddy, Load Forecasting by a novel technique using ANN, ARPN,VOL. 3, NO. 2,, pp:19-25, APRIL 2008

[6]. Nima Amjady, Day-Ahead Price Forecasting of Electricity Markets by a New Fuzzy Neural Network, IEEE Transactions On Power Systems, Vol. 21, NO. 2, pp:887-896,May 2006

[7]. Jan G. De Gooijer, Rob J. Hyndman, A look into the future of ‘electricity price forecasting’ , International Journal of Forecasting, Vol 30, Issue 4, Elsevier, pp:1030-1081, 2006

[8]. Charles C. Holt, Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages, International Journal of Forecasting, Vol 20, Issue 1, Elsevier, March 2004

[9]. Francisco J. Nogales, Javier Contreras, Antonio J. Conejo, Rosario Espínola, Forecasting Next-Day Electricity Prices by Time Series Models, IEEE Transactions On Power Systems, Vol. 17, No. 2,pp:342-348, May 2002

[10]. B.R. Szkuta, L.A. Sanabria, T.S. Dillon, Electricity Price Short-Term Forecasting Using Artificial Neural Networks, IEEE Transactions on Power Systems, Vol. 14, No. 3, pp:851-857, August 1999

Cite this article as :

Sindhuja T, Sakhi Chawda, Parimala Kanaga Devan Kailasam, "Vegetable Price Prediction using ARIMA", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 4, pp. 176-183, July-August 2021. Available at doi : https://doi.org/10.32628/CSEIT217440
Journal URL : https://ijsrcseit.com/CSEIT217440