RESEARCH REVIEW JOURNALS
Volume-03
Issue-08
August-2018

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Statistical Analysis of Food Grain Prices in Karnataka

[*1]Satyanarayana; [2]Swathi & [3]Ismail B.

[*1]*Research Scholar, Department of Statistics, Mangalore University, Mangalagangothri , Karnataka (India)*
[2]*Department of Statistics, Mangalore University, Mangalagangothri , Karnataka (India)*
[3]*Professor, Department of Statistics, Mangalore University, Mangalagangothri , Karnataka (India)*

ARTICLE DETAILS

Article History
Published Online: 07 August 2018

Keywords
ARIMA, Multi-Layer Perceptrons (MLP), Extreme Learning Machine (ELM)

[*]Corresponding Author
Email: sathya1301[at]gmail.com

ABSTRACT

Paddy and Wheat are the major food crops in Karnataka. Paddy is grown in 27 districts of Karnataka and out of which 14 districts are under high productivity group. Rice is a part of paddy. Wheat is grown in a larger area than any other crops in the world. The percentage share of production of wheat in total production by Karnataka is 0.31 %. Farmer's decision making on acreage under paddy and wheat depends on the future prices to be realized during the harvest period. Hence this paper presents different methodology like traditional time series method and neural network method to forecast the prices of rice and wheat in Karnataka. We compare the forecast accuracy of these approaches using accuracy measures like Root mean square error, mean absolute error and mean absolute percentage error. The identification of the best forecasting model would help the producers, consumers as well as suppliers in taking appropriate decisions. In case of rice price, ARIMA model was found to be the best forecasting model. In the wheat crop, the MLP model was found to be the best forecasting model. These models were used to forecast prices for next 12 months.

1. Introduction

Agriculture is the backbone of Indian economy. The major agricultural products can be broadly grouped into foods, fibers, fuels, and raw materials. Classes of foods include food grains, vegetables, fruits, oils, and meat. Over the one-third of the world's workers are employed in agriculture. Farming is a major source of income for many people in India. Paddy, ragi, and wheat are the major food grains are used in Karnataka.

Paddy is an important food crop in India and second most in the world. About 35% of the net cropped area under paddy and about 60% of the farmers cultivate paddy every year. Paddy becomes rice after the removal of husk by threshing. Rice is a part of paddy. Rice is India's predominant crop and is the staple food of the people of the eastern and southern part of the country. Rice is grown in Karnataka in 27 districts. Out of which 14 districts are under high productivity group. Nearly 54% rice area in Karnataka is concentrated in high productivity group and high productivity zone accounts for about 65% of total rice production in the State. It is mainly grown in Koppal, Davanagere, Bangalore, Mandya, and Mysore. It is mostly concentrated in the river valleys, deltas and low-lying coastal areas. Karnataka stands in 10th place in the production states of Karnataka. Karnataka accounts for more than 3% of total rice production in India. To meet the demand of increasing population and to maintain self-sufficiency IR8 is the high yielding rice variety developed and it saved many people's lives who are suffered from starvation.

Wheat is grown in larger areas than any other food crops in the world. Wheat can be grown as winter and spring crop. Wheat is a grass widely cultivated for its seed, a cereal grain which is a worldwide staple food. Wheat is the leading source of vegetable protein in daily human food. Wheat has higher protein content than other major cereals like maize and rice. With rice, wheat is the world's most favored staple food. It is less grown in Karnataka. The percentage share of production of wheat in total production by Karnataka is 0.31%. it is grown in Bijapur, Belgaum, Bagalkot, Bellary, Raichur and Koppal.

Agricultural price forecasts are an integral component of trade and policy analysis. As the prices of food grains directly influence the real income of consumers and it also affects the consumers' access to food. For the purpose of this study, two such food grains are selected i.e. rice and wheat because of huge demand and price fluctuation.

2. Methodology

A basic assumption in any time series analysis/modeling that some aspects of past pattern will continue to remain the future. Also under this setup, often the time series process is assumed to be based on past values of the main variable but on the explanatory variable which may affect variable. There are various objectives for studying time series. They include understanding and description of the generating mechanism, the forecasting of future values, and optimal control of a system. The intrinsic nature of a time series is that its observations are dependent or uncorrelated and the order of observation is therefore important. Hence, statistical procedures and techniques that rely upon independents assumptions are no longer applicable and different methods are needed. The body of statistical methodology available for analyzing time series is referred to as Time Series Analysis.

2.1 Testing for the presence of trend component

Mann-Kendall trend test:

The Mann-Kendall trend test is a nonparametric test used to identify a trend in a series, even if there is a seasonal component in the series.

H_0: there is no trend in the series

H_1: there is monotonic trend in the series.

The Mann-Kendall tests are based on the calculation of Kendall's tau measure of association between two samples, which is itself based on the ranks with the samples. The computations assume that the observations are independent.

2.2 Testing for the presence of seasonality component

H_0 : time series is free from seasonal variation
H_1: time series contains seasonal variation

Test statistic is

$$\chi_0^2 = \frac{12\sum_{j=1}^{D}(M_j - \frac{C(D-1)}{2})^2}{CD(D+1)} \sim \chi^2_{(D-1)}$$

D-seasonality periods, C-total number of years, M_j-sum of the ranks for the j^{th} period. If chi-square calculated is more than chi-square table value we reject H0 and conclude that there is seasonal variation is there in the data.

2.3 Autoregressive integrated moving average process

Let { X_t, tϵI } denotes a nonstationary time series, non stationary due to trend component. Let {ϵ_t, t =±1,±2,….} is a sequence of white noise. Then { X_t, tϵI } is said to follow autoregressive integrated moving average process if it has the representation

$$\Phi(B)\,(1-B)^d X_t = \theta(B)\,\epsilon_t$$

Where $\Phi(B) = 1 - \beta_1 B - \beta_2 B^2 - \ldots\ldots - \beta_p B^p$

$\theta(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \ldots\ldots - \alpha_q B^q$

$\alpha_1,\alpha_2,\ldots\alpha_q$ are MA parameters, $\beta_1,\beta_2,\ldots,\beta_p$ are AR parameters and d is the difference required to make given time series data to stationary time series. This model is also known as Box-Jenkins model.

2.3.1 Identification

The foremost step in the process of modeling is to check for the stationarity of the series, as the estimation procedures are available only for stationary series. A cursory look at the graph of the data and structure of autocorrelation and partial autocorrelation coefficients may provide clues for the presence of stationarity.

The next step in the identification process is to find the initial values for the orders of seasonal and non-seasonal parameters P, Q and p,q and it is obtained by looking for significant autocorrelation and partial autocorrelation coefficients.

Selection of stationary ARMA models:

The choice of the appropriate (p,q) values of the ARMA model for the stationary series is carried out on the grounds of its characteristics, that is, the mean, the ACF and the PACF.

Process	ACF	PACF
AR(p)	Infinite: exponential and/or sine-cosine wave decay	Finite: cut off at lag p
MA(q)	Finite: cut off at lag q	Infinite: exponential and/or sine-cosine wave decay
ARMA(p,q)	Infinite: exponential and/or sine-cosine wave decay	Infinite: exponential and/or sine-cosine wave decay

2.3.2 Estimation

At the identification stage, one or more models are tentatively chosen that seem to provide a statistically adequate representation of the available data. Then we attempt to obtain precise estimates of parameters of the model by least squares as advocated by Box and Jenkins. Standard computer packages like R languages etc.. are available for finding the estimate of relevant Parameters Using Iterative Procedures.

2.3.3 Diagnostics

Different models can be obtained for various combinations of AR and MA individually and collectively .Time series model building is an iterative procedure. It starts with model identification and estimation of parameter. After parameter estimation, we have to assess model adequacy by checking whether the q residuals are estimates of these unobserved white noise ϵ_t's. Hence, model diagnostic checking is accomplished through a careful analysis of the residual series.$\widehat{\epsilon_t}\}$.

2.3.4 Forecasting:

One of the most important objectives in the analysis of a time series is to forecast its future values. In forecasting, our objective is to produce and optimum forecast that has no error or as little error as possible, which leads us to the minimum mean square error forecast. This forecast will produce an optimum future value with the minimum error in terms of the mean square error criterion. Some of these measures are as follows

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

where X_{obs} is observed values and X_{model} is modelled values at time/place *i*.

$$MAE = \frac{\sum_{i=1}^{N}|x_i - \widehat{x_i}|}{N}$$

$\{x_i\}$ is the actual observations time series , $\{\hat{x}_i\}$ is the estimated or forecasted time series and N is the number of non-missing data points

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

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

Akaike Information Criteria:

The Akaike information criterion (AIC) is a measure of the relative quality of a statistical model for a given set of data. AIC provides a means for model selection.

$$AIC = -2\ln(L) + 2p,$$

where ln(L) denotes the fitted log likelihood and p the number of parameters. Given the set of candidate models for the data, the preferred model is the one with the minimum AIC value.

2.4 Neural Network

A neural network is a biologically inspired nonlinear parallel computing paradigm for information processing and exploratory analysis having a distinct ordering among the sets of neurons arranged as input and output layers with zero or more processing layers that are interconnected by signal channels and fine tuned by training algorithm. A neural network is a set of connected input/output units in which each connection has a weight associated with it. The weights are adjusted during the learning phase to help the network predict the correct class label of the input tuples. Neural network learning is also referred to as connectionist learning due to the connections between units. Neural networks involve long training times and are therefore more suitable for applications where this is feasible. They require a number of parameters that are typically best determined empirically such as the network topology or "structure."

Advantages of neural networks, however, include their high tolerance of noisy data as well as their ability to classify patterns on which they have not been trained. They can be used when you may have little knowledge of the relationships between attributes and classes. They are well suited for continuous-valued inputs and outputs, unlike most decision tree algorithms. Neural network algorithms are inherently parallel; parallelization techniques can be used to speed up the computation process. In addition, several techniques have been recently developed for rule extraction from trained neural networks.

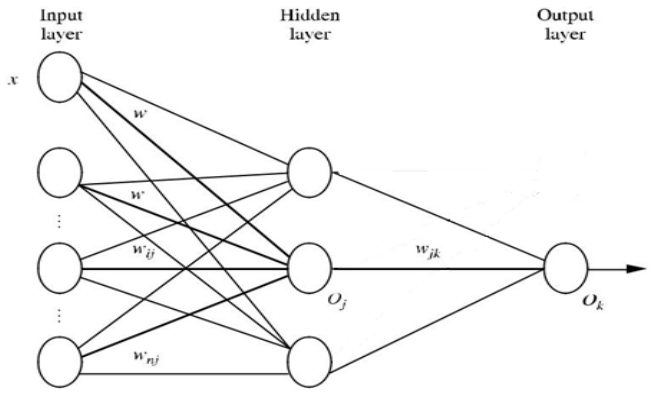
Components of neural network

1. Network architecture: learns relationship between inputs and output values
2. Activation and transfer function: transforms input values using synaptive strengths to a neuron
3. Synaptic weights: they are pre specified by user and is obtained from previous runs of the model.
4. Training algorithm: simulates learning algorithm
5. Training and testing sets: builds intelligence to solve practical problems

2.4.1 Multilayer Perceptron Method (MLP)

A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer. A multilayer feed forward neural network is an interconnections of perceptrons in which data and calculations flow in a single directions, from the input data to the outputs. The number of layers in a neural network is the number of layers of perceptrons. The simplest neural network is one with a single input layer and an output layer of perceptrons.

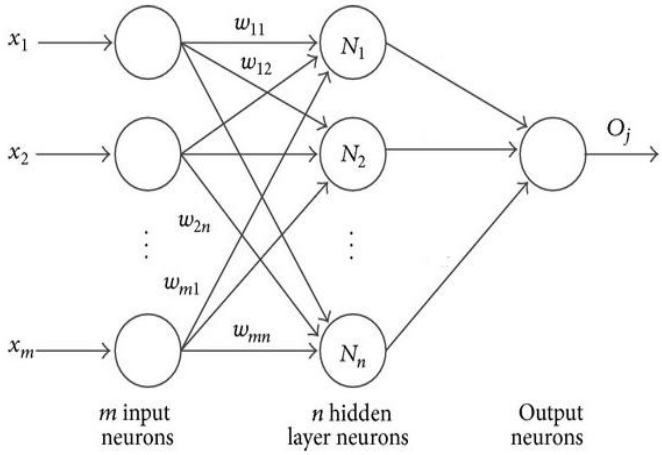
The next most complicated neural network is one with two layers. This extra layer is referred to as a hidden layer. In general ,there is no restriction on the number of hidden layers. The back propagation algorithm performs learning on a multilayer feed-forward neural network. It iteratively learns a set of weights for prediction of the class label of tuples. However, increases the number of perceptrons increases the number of weights that must be estimated in the network, which in turn increases the execution time for the network. Instead of increasing the number of perceptrons in the hidden layer to improve accuracy, it is sometimes better to add additional hidden layers, which typically reduce both the total number of network weights and the computational time.



Each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of "neuronlike" units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used.

2.4.2 Extreme Learning Machine (ELM)

Extreme learning machine are feed-forward neural networks compression and feature learning with a single layer or multiple layers of hidden nodes, where the parameters of hidden nodes need not be tuned. These hidden nodes can be randomly assigned and never updated or can be inherited from their ancestors without being changed. In most cases, the output weights of hidden nodes are usually learned in a single step, which essentially amounts to learning a linear model.



Back propagation Algorithm

Propagation learns by iteratively processing a data set of training tuples, comparing the network's prediction for each tuple with the actual known target value. For each training tuple, the weights are modified so as to minimize the mean-squared error between the network's prediction and the actual target value.

To compute the net input to the unit, each input connected to the unit is multiplied by its corresponding weight, and this is summed. Given a unit, *j* in a hidden or output layer, the net input, I_j , to unit *j* is

$$I_j = \sum_i w_{ij} O_i + \theta_j$$

where w_{ij} is the weight of the connection from unit i in the previous layer to unit j. O_j is the output of unit i from the previous layer and j is the bias of the unit.

Backpropagate the error: The error is propagated backward by updating the weights and biases to reflect the error of the network's prediction. For a unit *j* in the output layer, the error Err_j is computed by

$$Err_j = O_j(1 - O_j)(T_j - O_j)$$

3. Results and Discussions

The dataset which we use for analysis is collected from agriculture prices in India, Publication: Economic statistics, Ministry of agriculture in India. The analysis was done on 10 years data for rice and wheat from 2008-2017. The data from the year 2008-2016 is taken as training data and 2017 year data is taken as testing data. The analysis is carried out using R software.

Table1: Represent the descriptive statistics of rice price

Mean	**2089.194**
Standard deviation	391.7576
Skewness	-0.14277
Kurtosis	1.9408

The average value of rice price is Rs 2089 with standard deviation 391. Here kurtosis value is less than 3 so it is less tailed and skewness value indicates that negatively skewed.

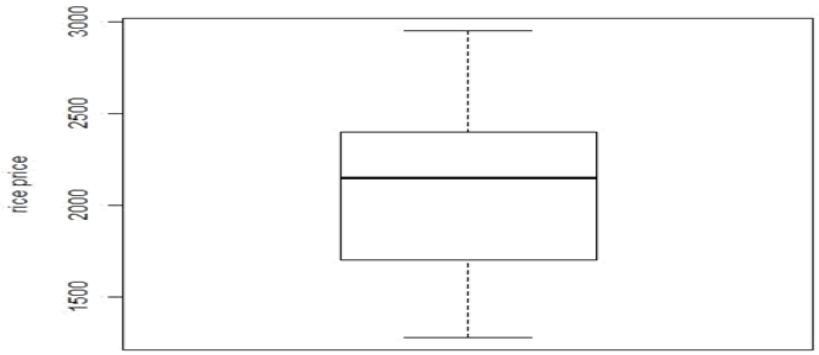


Fig 1 : Box plot of the rice price data

From the above plot, upper line and lower line shows the maximum and minimum value of data are nearly equals to 2900 and 1200 respectively and middle line of the box shows median value is nearly equal to 2150, upper line and lower line of the box shows third quartile and first quartile value of the data are nearly 2400 and 1700 respectively. There is no outlier in the data. The data is negatively skewed.

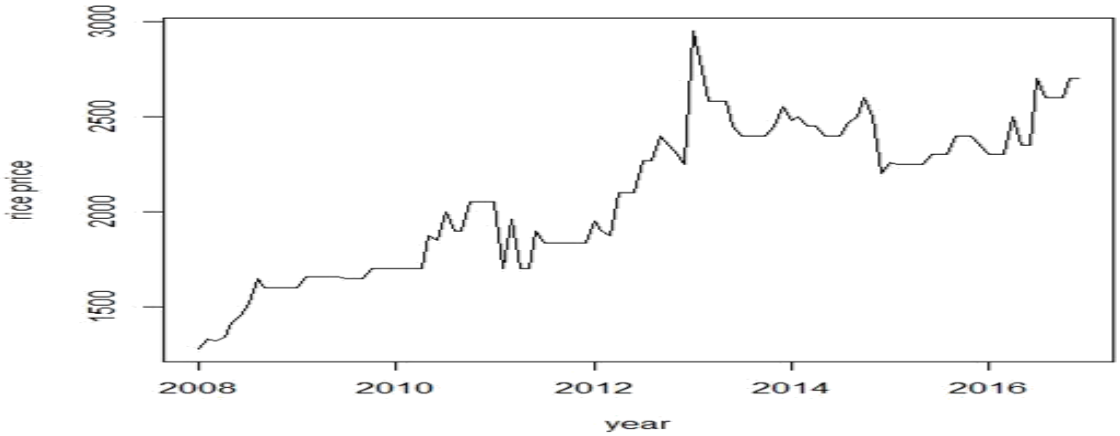


Figure 2: Time profile for rice price of Karnataka over the year 2008-2017

Rank-sum test

H0: there is no seasonal variation in the data
H1: there is seasonal variation in the data

Chi-square calculated value is 13.8974 and Chi-square critical value is 19.675

Since calculated value of chi square is less than critical value of chi square we accept H0 and conclude that there is no seasonal variation in the data

Mankendall test – test for trend

H0: There is no trend in the series
H1: There is a trend in the series

Calculated value = 0.703, 2-sided and p value = < 2.22e-16

As the computed p-value is less than the significance level alpha=0.05, we reject the null hypothesis and conclude that there is trend in the given data.

Variance difference method: The data consider for the analysis has only trend component. Therefore, we carried out variance difference method to make the series stationary Variance of given series is 153474, Variance of first difference series is 13908.14 and the Variance of second difference series is 35414.77. Since variance of second difference series is more than the variance of first difference series, first difference series is stationary.

Based on the ACF and PACF plot we fitted the different ARIMA model with different order. We select the best model for which AIC value is minimum and p value maximum.

Table 2: Represents the summary of the different fitted model

(p,d,q)	AIC	Ljung-Box p-value
(1,1,0)	1321.92	0.9167
(1,1,1)	1320.16	0.9698
(0,1,1)	1318.60	0.9793
(2,1,0)	1318.46	0.9791
(2,1,1)	1320.37	0.9716

From the above table we observe that ARIMA (2,1,0) model has minimum AIC value and maximum Ljung-Box p-value.

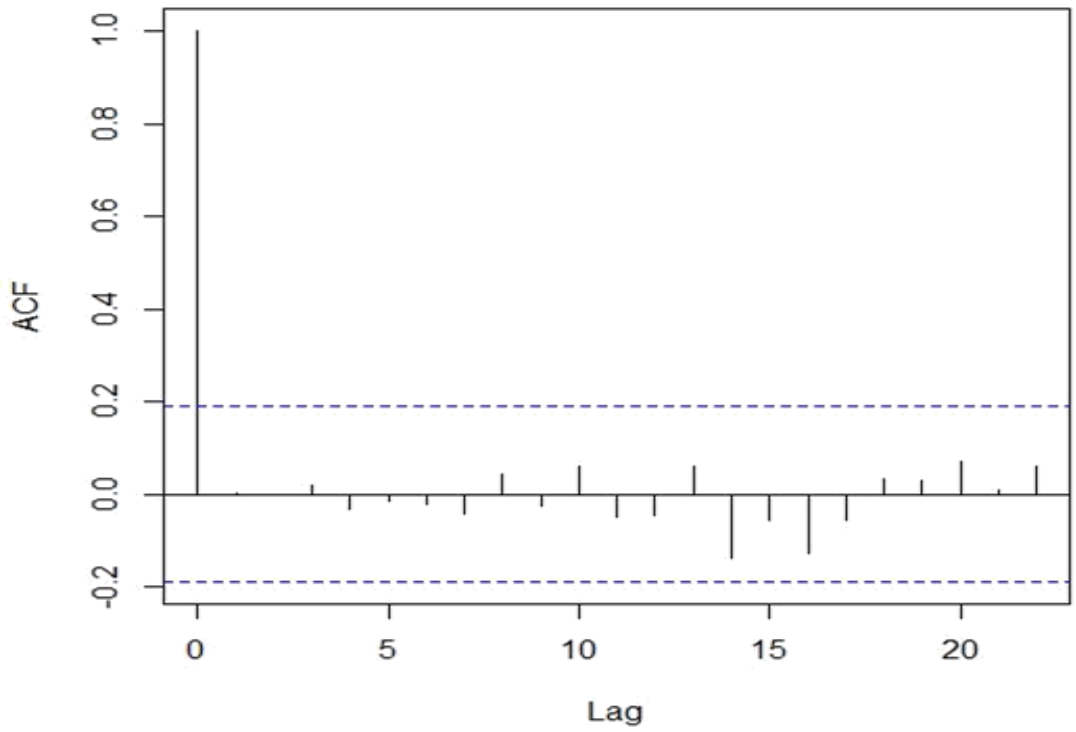


Figure 3: Residual series for ARIMA (2,1,0)

Figure 3 represents the ACF of residual series for fitted ARIMA (2,1,0). We can observe that there are insignificant ACF, which indicates that residual series are uncorrelated. It behaves as white noise.

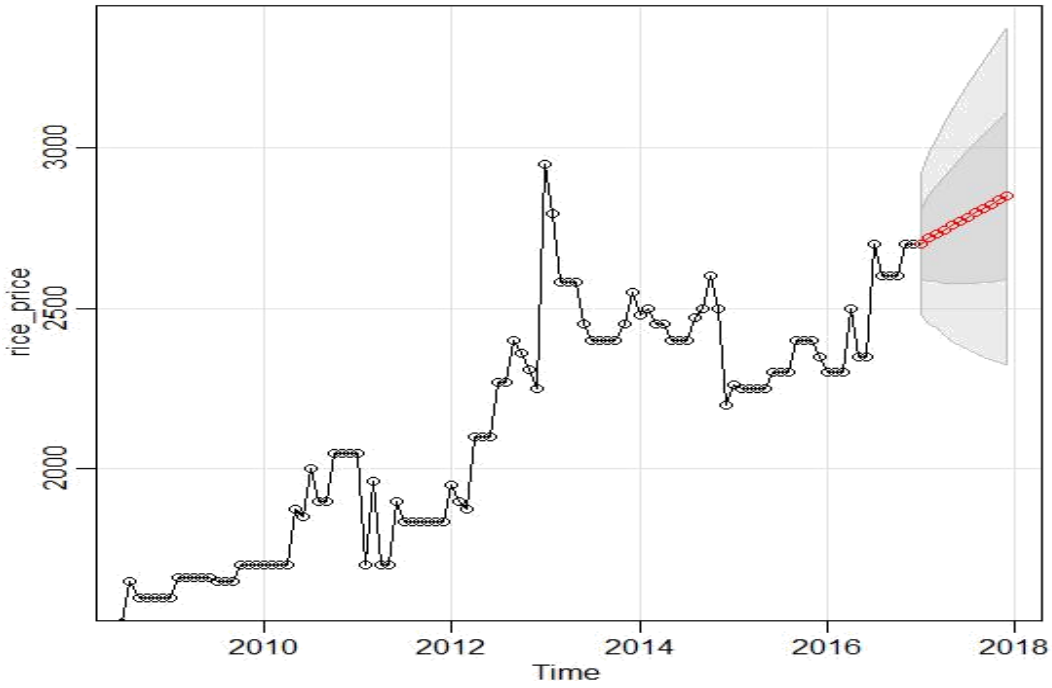


Figure 4: Forecast from ARIMA (2,1,0) model

The constructed MLP has 5 hidden nodes and 20 repetitions. Forecast combined using the median operator.

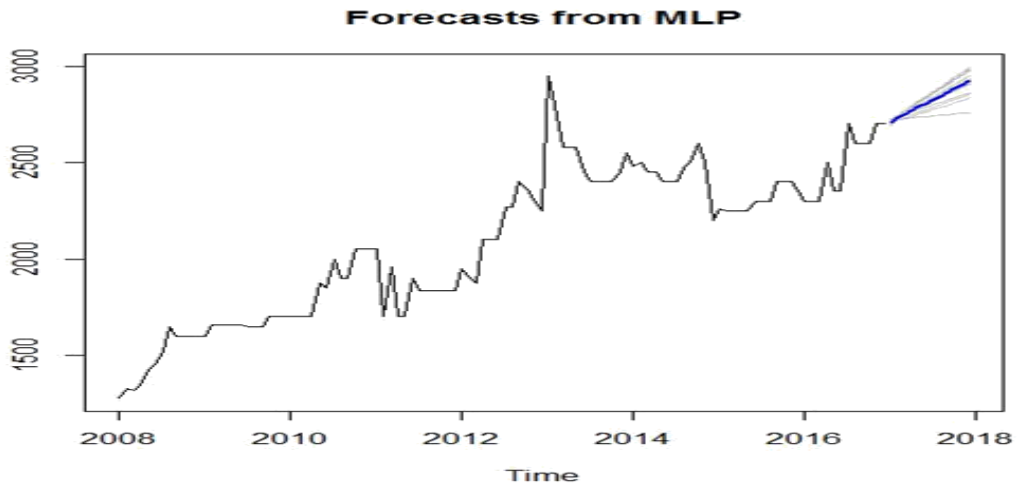


Figure 5: Forecast from MLP network for rice

The constructed ELM has 100 hidden nodes and 20 repetitions. Forecast combined using the median operator. Output weight estimation using: lasso.

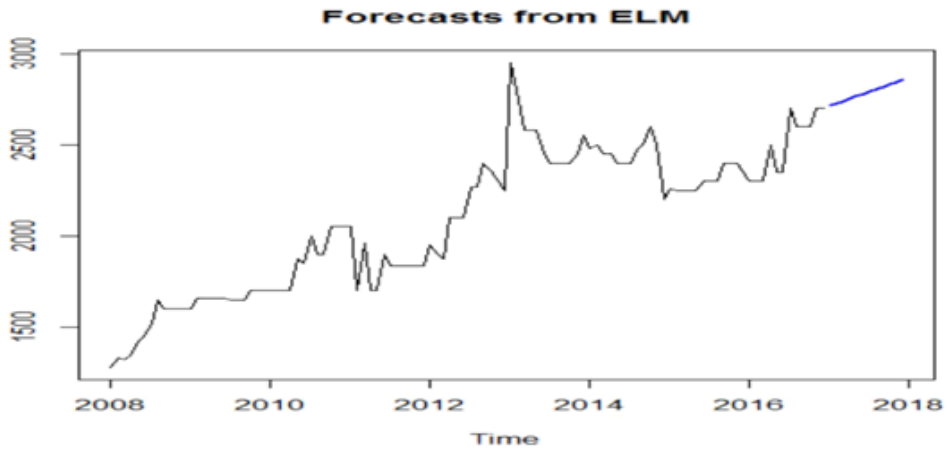


Figure 6: Forecast from ELM network for rice

Table 3: Represents the accuracy measures for different models

	RMSE	MAE	MAPE
ARIMA(2,1,0)	**62.7191**	**56.2985**	**2.0646**
MLP	110.559	98.6014	3.6128
ELM	70.1565	64.5952	2.3695

From the table 5, since the value of RMSE , MAE, MAPE is less for ARIMA(2,1,0) model compared to MLP and ELM model, ARIMA(2,1,0) model is best model for forecasting the rice price.

Table 4: Represent the descriptive statistics of wheat price

Mean	2098.611
Standard deviation	427.0415
Skewness	0.1242
Kurtosis	1.85567

The average price of wheat is Rs 2098 with standard deviation of Rs 427. Here kurtosis value is more than 3 so that the distribution is leptokurtic and the skewnes value indicates that it is positively skewed.

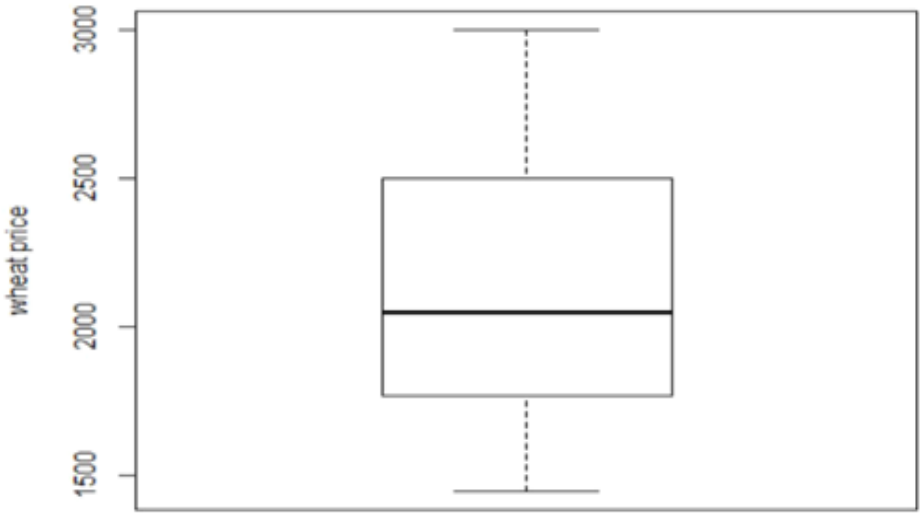


Figure 7: box plot of wheat price data

From the above plot, upper line and lower line shows that maximum and minimum value of data is nearly equals to Rs 3000 and Rs 1450 respectively and middle line of the box shows that median value is nearly equal to Rs 2050, upper line and lower line of the box shows third quartile and first quartile value of the data are nearly Rs 2500 and Rs 1770 respectively. There is no outlier in the data.

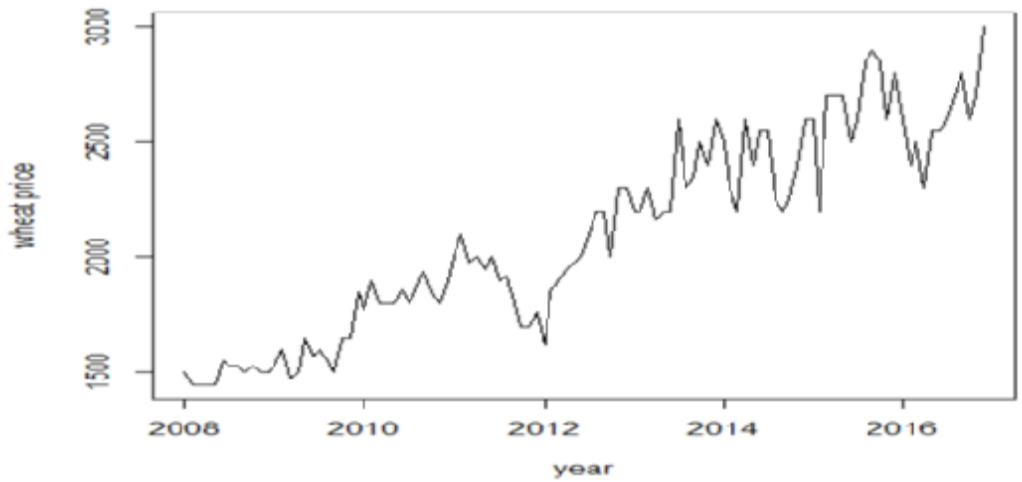


Figure 8: time profile of Wheat price in Karnataka over the year 2008-2017

Rank-sum test

H0: there is no seasonal variation in the data
H1: there is seasonal variation in the data

Chi-square calculated value is 19.0174. Chi-square critical value is 19.6754

Since calculated value of chi square is less than critical value of chi square we accept H0 and conclude that there is no seasonal variation in the data.

Mankendall test – test for trend

H0: There is no trend in the series
H1: There is a trend in the series

Calculated value = 0.804, 2-sided p value = < 2.22e-16

As the computed p-value is less than the significance level alpha=0.05, we reject the null hypothesis and conclude that there is trend in the given data.

Variance difference method: The data consider for the analysis has only trend component. Therefore We carried out variance difference method to make the given series stationary. Variance of given time series is 182364.4, Variance of first difference series is 21931.81 and Variance of second difference series is 57737.57. Since variance of second difference series is more than the variance of first difference series, first difference series is stationary.

Based on the ACF and PACF plot we fitted the different arima model with different order. We select the best model for which AIC value is minimum and p value maximum.

Table 5 : Represents the summary of the different fitted model

(p,d,q)	AIC	Ljung-Box p-value
(0,1,1)	1464.61	0.9595
(1,1,0)	1497.11	0.3351
(1,1,1)	1463.82	0.8875
(2,1,0)	1474.21	0.7494
(2,1,1)	1462.76	0.9609

From the above table, we observe that for ARIMA (2,1,1) model, AIC value is minimum and Ljung-Box p-value is maximum.

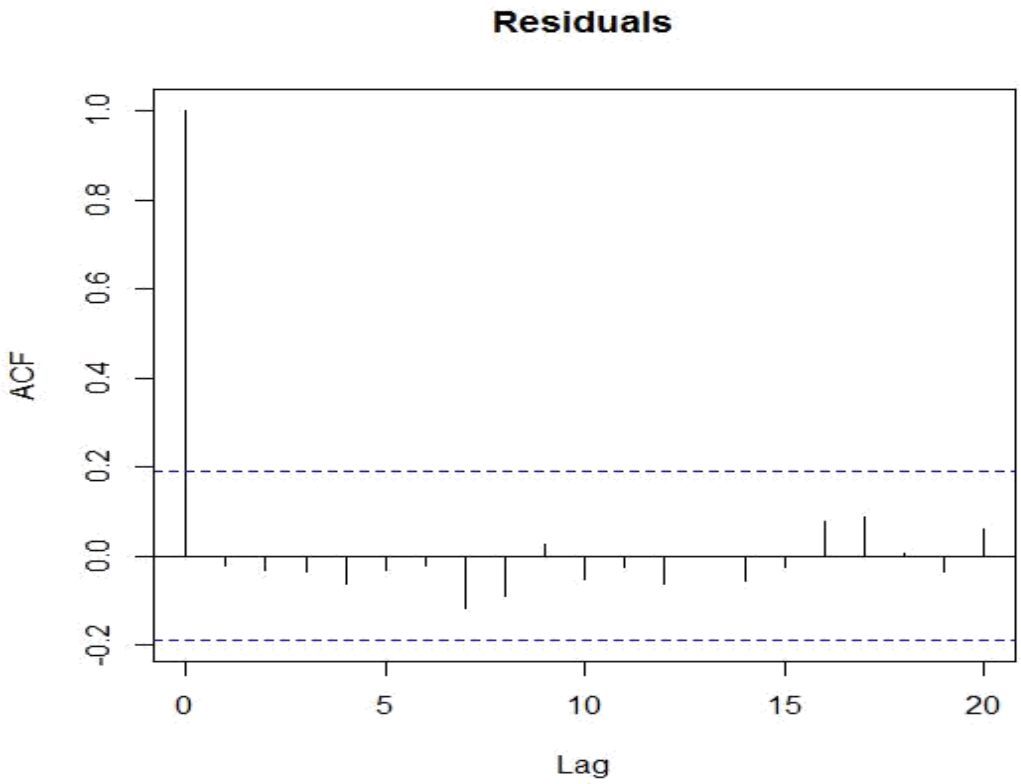


Figure 9: Residual series for ARIMA (2,1,1)

Figure 9 represents the residual series for fitted ARIMA (2,1,1).We can observe that there are insignificant ACF, which indicates that residual series are uncorrelated. It behaves as white noise.

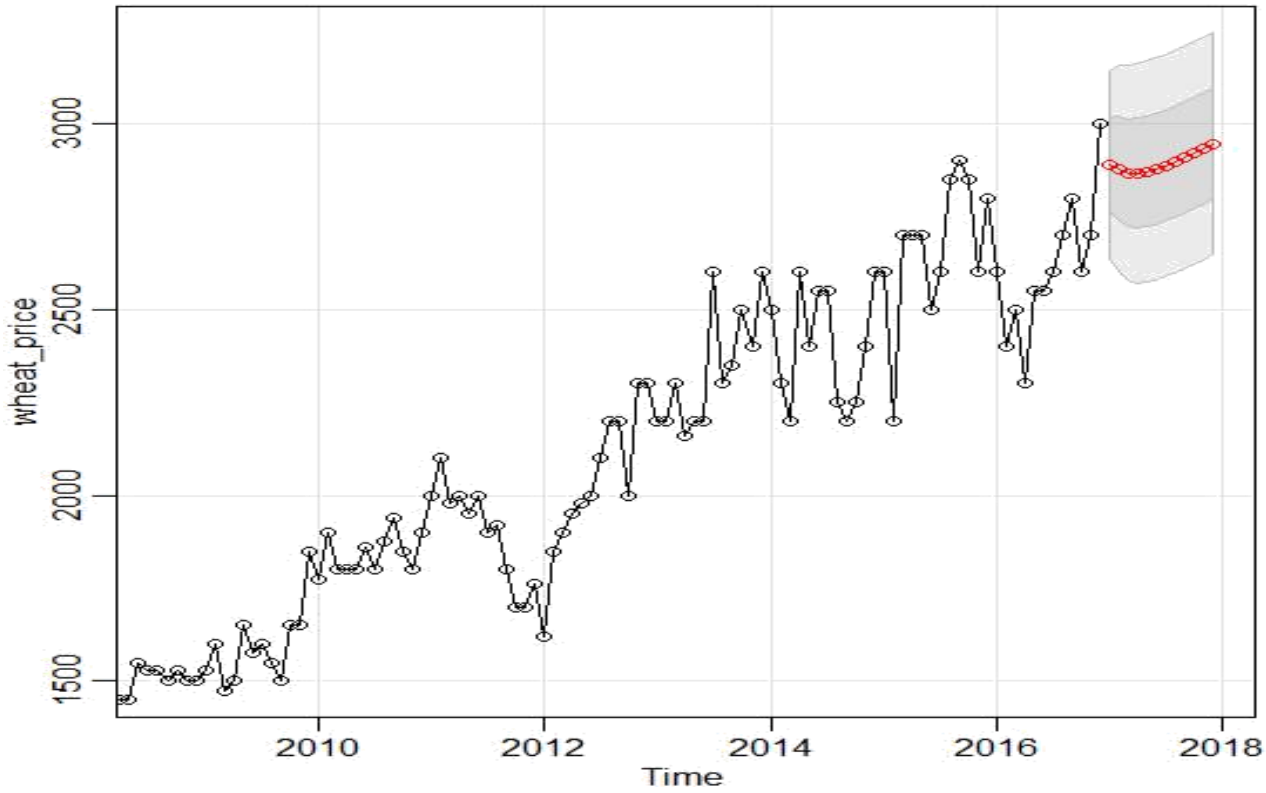


Figure 10 : Forecast from ARIMA (2,1,1)

The constructed MLP model has 6 hidden nodes and 20 repetitions. Forecast combined using the median operator.

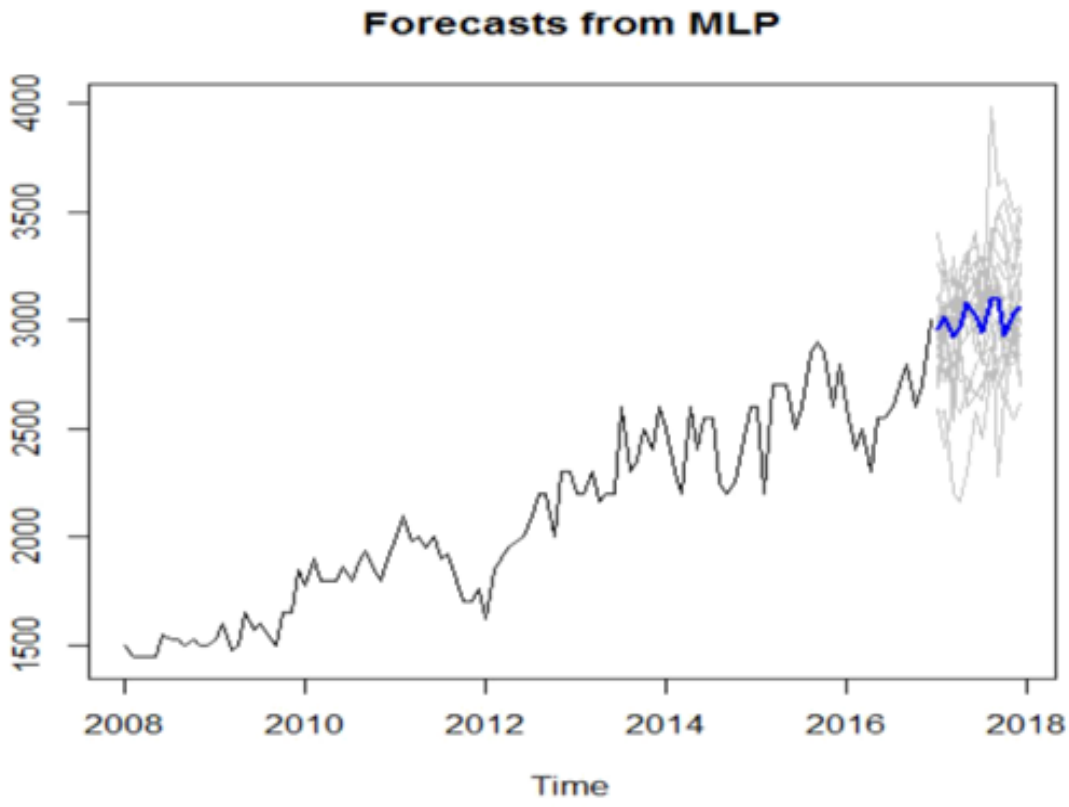


Figure 11 : Forecast from MLP

The constructed has 95 hidden nodes and 20 repetitions. Forecast combined using the median operator. Output weight estimation using lasso.

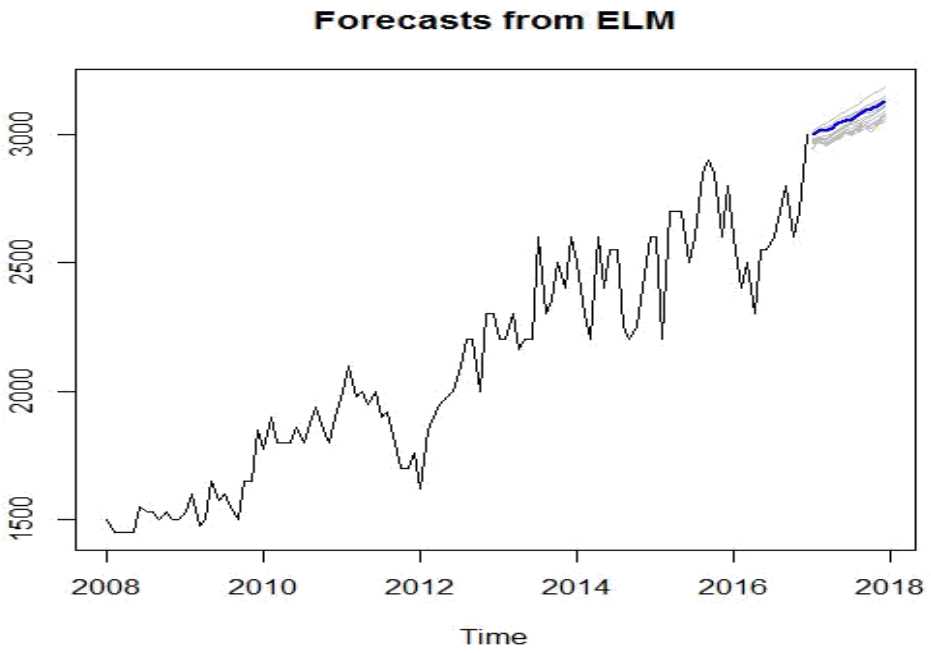


Figure 12 : Forecast from ELM

Table 6: Represents the comparison of ARIMA forecast, MLP forecast and ELM forecast with actual value

Year 2017	Original	Forecast from ARIMA(2,1,1)	Forecast from MLP	Forecast from ELM
January	2900	2724.859	2961.021	3014.903
February	2950	2732.142	3014.295	3029.807
March	2850	2781.802	2922.598	3044.711
April	3000	2791.711	2966.080	3059.615
May	2900	2788.328	3082.674	3074.519
June	2800	2798.211	3026.522	3089.423
July	3150	2815.169	2940.526	3104.326
August	2950	2826.014	3092.602	3119.230
September	2550	2834.427	2805.905	3134.134
October	2550	2844.970	2850.179	3149.038
November	2800	2856.726	2902.918	3163.942
December	2800	2867.451	2923.001	3178.846

Table 7: Represents the accuracy measures for different models

	RMSE	MAE	MAPE
ARIMA(2,1,1)	191.863	162.105	5.721
MLP	**178.485**	**131.680**	**4.783**
ELM	313.850	254.487	9.329

From the above table it is clear that on the value of RMSE, MAE, MAPE is less for MLP model compared to ARIMA (2, 1, 1) and ELM model, MLP model is the best model for forecasting the wheat price.

3. Conclusion

The food grain crops prices not only depend on domestic demand and supply conditions but also determined by global supply and demand situations. In the recent year's fluctuation in prices increased significantly with a lot of uncertainty in profitability of growing food crops like rice and wheat. Under this scenario, forecasting future prices will help the farmers to facilitate informed decisions to choose the profitable crops before sowing the crops. This paper used historical monthly prices of rice and wheat in Karnataka to forecast future prices for the next 12 months using the best model among ARIMA, MLP, and ELM model. We compare the forecast accuracy of these approaches using accuracy measures like Root mean

square error, mean absolute error and mean absolute percentage error. The identification of the best forecasting model would help the producers, consumers as well as suppliers in taking appropriate decisions. From the study, we conclude that in case of rice price, ARIMA model was found to be the best forecasting model. In the wheat crop, the MLP model was found to be the best forecasting model. The validity of the forecasted value for the year 2017 is checked with the available data.

Finally, we forecasted the future price of these crops using the best model for the next 12 months. This forecast is based on past data and the selected model and that the actual market price may not turn out to be the same as forecast.

References

1. Brockwell P.J and Davis R.S(2002): Introduction to Time Series and Forecasting, 2nd Ed., Springer.
2. Chatfield.C(1980):'An Introduction to the Analysis of Time Series', 2nd Ed., Chapman and hall, Londen.
3. Kendall M.G. and Ord J.K. (1990): Time Series, 3rd Ed, Edward Amold.
4. Jiawei Han, Micheline Kamber(2002): Data Mining-Concepts and Techniques, Morgan Kaufman Publishers, U.S.A.
5. Simon Haykin (2012): Neuarl Networks and Learning Machines, 3rd edition, PHI Learning Private Limited, New Delhi
6. Box GEP and Jenkins G.M(1976): Time Series Analysis: Forecasting and Control, Holden-day, San Franscisco.
7. Darekar A.S, Pokharkar V.G. and Yadav D.B (2016): Onion Price Forecasting In Yeola Market of Western Maharashtra Using ARIMA Technique, International Journal of Advanced Biological Research, 6(04): 551-552.
8. Jalika V.N. and Patil B.L(2015): Onion price forecasting in Hubli market of Northern Karnataka using ARIMA technique, Karnataka ,Journal Agricultural Science , 28(2): 228-231.
9. Pravin A, Singh D.R and Sivaramane N (2005) : An Application of Box-Jenkins Approach for Forecasting Copra Wholesale Price Series, Indian Society of Agricultural Statistics, 59: 32-47.
10. Makridakis S and Hibbon M (1979): Accuracy of forecasting: An empirical investigation, Journal of Royal Statistical Society of America. 41(2): 97-145.
11. Sivapathasundaram V and C. Bogahawatte (2012) : Forecasting of Paddy Production in SriLanka: A Time Series Analysis using ARIMA Model, Tropical Agricultural Research, Vol. 24 (1): 21 - 30

