Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Forecasting Prices of Soya bean Oil Using Statistical Models

Singh Abhishek

Dept. of Farm Engineering ; Institute of Agricultural Sciences; Banaras Hindu University, Varanasi UP,India221005

ABSTRACT

Soybean is an important crop of India and its production is affected by weather vagaries such as unpredictable rains and rise in temperatures. Soybean oil is used in Indian kitchens for its health benefits and nutritional values. The prices of soy oil show huge fluctuations. In this paper times series namely ARIMA (Autoregressive Integrated Moving Average) methodology given by Box and Jenkins has been used for forecasting prices of Soy oil. This approach has been compared with ANN (Artificial Neural Network) methodology. The results showed that therelativeperformance varies acrossforecasting horizons and different methodsperform best for different forecasting horizons.

KEYWORDS: Forecasting; Feed forward network; ARIMA; ANN

***Corresponding author:**

Dr. Abhishek Singh,

Assistant Prof., Dept. Of Farm Engineering,

Institute of Agricultural Sciences,

Banaras Hindu University,

email;asbhu2006@gmail.com Mob no. 9451526775

INTRODUCTION

India is one of the major oilseeds producing country in the world. India has diverse agro-climatic conditions which are suitable for growing various oilseed crops and varieties. Nine major oilseeds crops which are grown in India are - Groundnut, Rapeseed Mustard, Sunflower, Sesamum, Soybean, Safflower, Castor, Linseed and Niger. Soybean oil is one of the important vegetable oil used in India. In India maximum production of soybean comes from the state of Madhya Pradesh, followed byMaharashtra and Rajasthan.Main season for Soy bean cultivation is the kharif season (sown in June and July and harvested in November and December). Due to unpredictability of the Indian monsoon yield of soy bean show much variability. The peak arrival of soybean takes place during October–November. Soybean prices follow domestic as well as international sentiments and display high volatility.Agricultural commodity marketing data, especially the price data are vital for any future agricultural development project because they can influence potential supply and demand, distribution channels of agricultural commodity and the economics of agriculture. So price forecasting is expected to reduce the uncertainty and risk in the agriculture commodity market and can be used to determine the quantity of food grains and food product consumed, and to identify and make appropriate and sustainable food grain policy for the government.

There are many methods for analyzing a time series but one of the most simple and bench mark method is that of Box and Jenkins[1]which is popularly known as ARIMA methodology. De Gooijeret.al[2].provided an excellent review of time series methods in forecasting. The problem with ARIMA methodology is that it assumes a linear structure and stationarity of the process of which a particular times series is a realization, which is often not fulfilled. To overcome this limitation of the ARIMA methodology, Artificial Neural Networks (ANN) has also been used to forecast the prices as shown by Kohzadi Nowrouz et al.[3], Tang et al.[4] and Zoua et al.[5] Artificial Neural Networks effectively covers both linear and non linear processes, stationary as well as non stationary time series. The tremendous success of the Artificial Neural Networks can be attributed to some of its distinct characters such as its power to model extremely complex function in particular the non linear functions. They can also handle the problem of parsimony in linear models.

MATERIAL AND METHODS

Monthly cash prices of Soybean oil Indore in from January 1992 to July 2010 are used to test the prediction power of the two approaches. Data are obtained from the Website (www.sopa.org). An ARIMA model was estimated using the SPSS 16.0 statistical package. The model was then used to forecast on its respective three month out-of-sample set.

In the case of the neural networks, the time series was divided into a training, testing, and a validation (out-of-sample) set. The out-of-sample period was identical to the ARIMA model. The models used for forecasting are described below:

Auto regressive integrated moving average (ARIMA) time series model: Introduced by Box and Jenkins, the ARIMA model has been one of the most popular approaches for forecasting. In an ARIMA model, the estimated value of a variable is supposed to be a linear combination of the past values and the past errors. Generally a non seasonal time series can be modeled as a combination of past values and errors, which can be denoted as ARIMA (p,d,q) which is expressed in the following form:

$$X_t = \theta_0 + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \ldots\ldots\ldots + \Phi_p X_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots\ldots\ldots - \theta\, e_{t-q}, \quad (1)$$

where X_t and e_t are the actual values and random error at time t, respectively, $\Phi_i (i = 1,2,\ldots\ldots,p)$ and $\theta_j (j = 1,2,\ldots\ldots,q)$ are model parameters, p and q are integers and often referred to as orders of autoregressive and moving average polynomials respectively. Random errors $e_t$are assumed to be independently and identically distributed with mean zero and the constant variance, σ_e^2. Basically this method has three phases: model identification, parameters estimation and diagnostic checking.

Artificial neural network (ANN) model: Neural Networks are simulated networks with interconnected simple processing neurons which aim to mimic the function of the brain central nervous system. Mcculloch and Pitts[6] for the first time proposed the idea of the artificial neural network (ANN) but because of the lack of computing facilities they were not in much use until the back propagation algorithm was discovered by Rumelhart et al[7]. Neural networks are good at input and output relationship modeling even for noisy data. The greatest advantage of a neural network is its ability to model complex non linear relationship without a priori assumptions of the nature of the relationship. Apart from this artificial neural networks can also be used for classification problems as was shown by Ripley[8]. The ann model performs a nonlinear functional mapping from the past observations ($X_{t-1}, X_{t-2,}\ldots\ldots\ldots, X_{t-p}$) to the future value X_t i.e.

$$X_t = f(X_{t-1}, X_{t-2,}\ldots\ldots\ldots, X_{t-p,} w) + e_t \quad (2)$$

where w is a vector of all parameters and f is a function determined by the network structure and connection weights.

Training of the neural network is essential factor for the success of the neural networks.Among the several learning algorithms available, back propagation has been the most popular and most widely implemented learning algorithm of all neural networks paradigms. The important task of the

ANN modeling for a time series is to choose an appropriate number of hidden nodes, q, as well as the dimensions of the input vector p (the lagged observations).

Criteria for comparing the prediction accuracy of arima and ann procedures: Different criteria were used to make comparisons between the forecasting ability of the ARIMAtime series models and the neural network models. The first criterion is the absolute mean error (AME). It is a measure of average error for each point forecast made by the two methods. AME is given by

$$AME = \left(1/T\right)\sum|P_t - A_t| \qquad (3)$$

The second criterion is the mean absolute percent error (MAPE). It is similar to AME except that the error is measured in percentage terms, and therefore allows comparisons in units which are different.

The third criterion is mean square error (MSE), which measures the overall performance of a model. The formula for MSE is

$$\text{MSE} = \left(1/T\right)\sum(P_t - A_t)^2 \qquad (4)$$

where$P_t$is the predicted value for time t, A_t is the actual value at time t and T is the number of predictions and the fourth criterion is RMSE which is the square root of MSE.

RESULTS AND DISCUSSION:

ARIMA model

For fitting the ARIMA model the three stages of modeling as suggested by Box and Jenkins namely, identification, estimation and diagnostic checking was undertaken. Identification was done after examining the auto correlation function and the partial autocorrelation function. After that estimation of the model was done by the least square method. In the diagnostic checking phase the model residual analysis was performed.

Figure 1 shows the time plot prices of the soybean oil in Indore. By the graph it can be inferred that the series is not stationary because the mean of the time series is increasing with the increase in time. So the time series is showing an increasing trend but to confirm this auto correlation function was calculated. Box and Jenkins suggest that the most autocorrelations examined is about one-fourth of the number of observations. In the present case 50 autocorrelations were calculated.

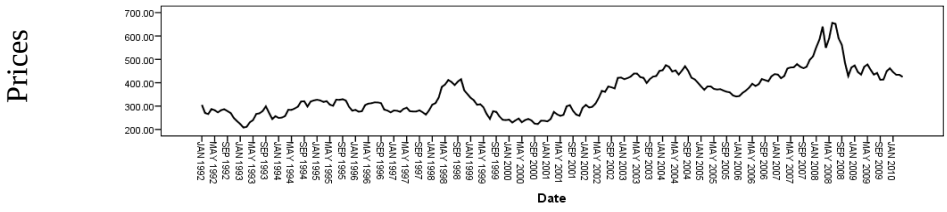

Fig 1: The time plot of prices of the Soybean oilin Indore

In the fig: 2 the auto correlation function of the time series shows that the series is not stationary because auto correlation coefficients do not cut off to statistical insignificance fairly quickly. All the first 50 autocorrelations are significantly different from zero at about the 5% level: all the first 50 spikes in the acf extend beyond the square brackets. The position of those brackets is based on Bartlett's approximation for the standard error of estimated autocorrelations. The brackets are placed about two standard errors above and below zero. To make the series stationary it was first differenced.

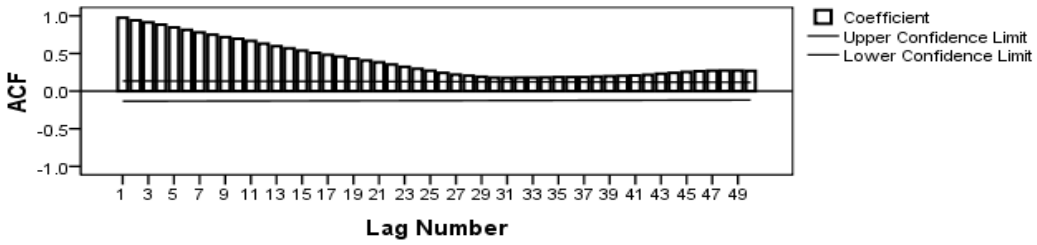

Fig 2: Autocorrelations at different lags of Soybean oil in Indore

Figure3 shows the time plot of the differenced series and it clearly depicts that the series has now become mean stationary. By looking at the variance of the series log transformation of the data was taken. The observations fluctuate around a fixed mean, and the variance is varying over time. However, the judgment about stationarity of the mean was withheld until the estimated acf and estimated AR coefficients were examined.

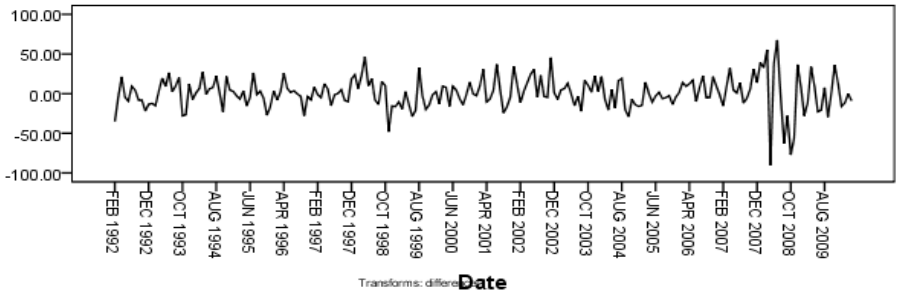

Fig 3: Time plot of the differenced series of Soybean Oil in Indore

In fig 4 and fig 5 autocorrelation function and partial autocorrelation function of the differenced series are shown. The autocorrelations decay to statistical insignificance rather quickly. It was concluded that the mean of the series is stationary

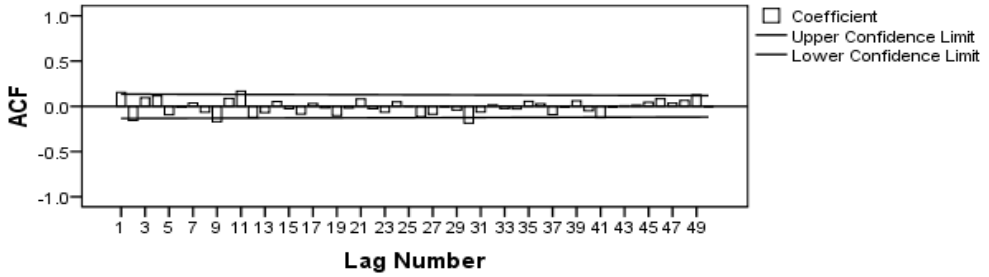

Fig 4: Autocorrelations at different lags of differenced time series of Soybean oil in Indore

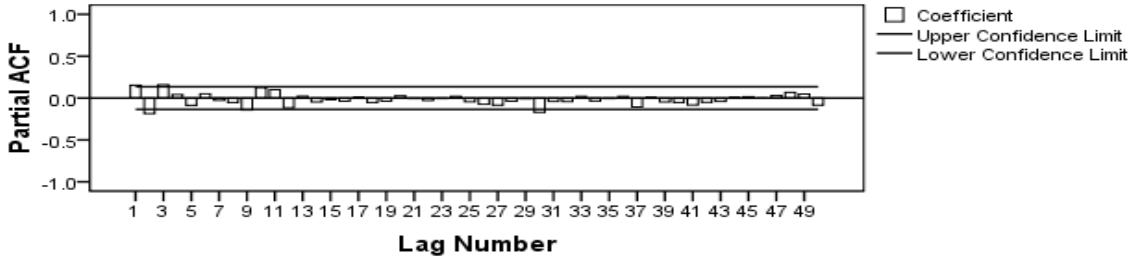

Fig 5: Partial Autocorrelations at different lags of differenced time series Soybean oil in Indore

Once the time series has become stationary using expert modeler option in SPSS, the ARIMA model was estimated. After going through these stages ARIMA (0,1,11) model was found to be the best among the family of ARIMA models . ARIMA Model parameters and model Fit statistics are given in Table-1

Table 1:ARIMA Model parameters and model Fit statistics for Soybean oil in Indore

	Estimate	SE	T	Sig	Model Fit Statistics	
Constant					Stationary R Squared	0.092
Difference	1				R Squared	0.957
MA					RMSE	19.241
Lag 1	-0.263	0.64	-4.116	0.00	MAPE	3.882
Lag 11	-0.216	0.64	-3.346	0.001	MAE	13.757
					Normalized BIC	5.963

This model satisfies the stationarity requirement $\theta_1 + \theta_{11} < 1.0$. Also θ_1 and θ_{11} are significantly different from zero at better than the 1% level since its absolute t-value of 4.116 at lag 1 and absolute t-value of 3.346 at lag 11 which is greater than2.0. Also R^2 value is 0.957 and RMSE ,MAPE, MAE, BIC is 19.241, 3.882, 13.757 and 5.963 respectively showing satisfactory model fitting.

At the diagnostic checking stage residual were examined and their autocorrelation coefficients were found to be non significant. (Fig-6) .Which shows that the model is satisfactory.

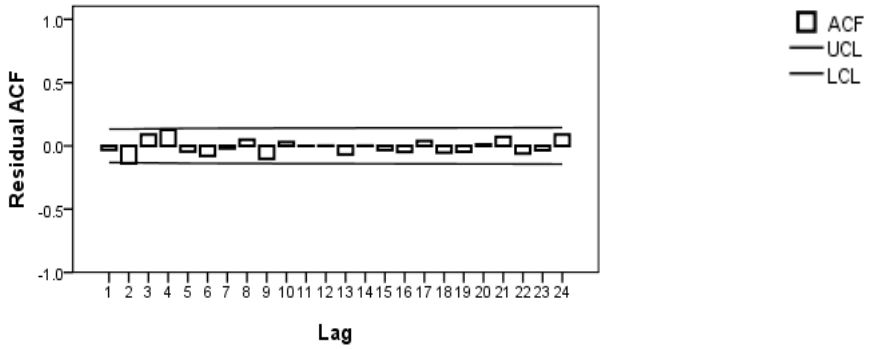

Fig 6: Residual auto and partial autocorrelations for Soybean oil in Indore

To determine if model is statistically adequate, the random shocksfor independence using the residuals from the estimated equation were tested. The residuals are estimates of the random shocks, and these shocks are assumed to be statistically independent. The estimated acf of the residuals were used to test whether the shocks are independent. With 150 residuals about 24 residual autocorrelations were examined. The residual acf appears in Figure 6. None of the residual autocorrelations has an absolute t-value exceeding the warning levels ie 1.25 at lags 1, 2, and 3 and 1.6 elsewhere. Because there is no dependence among the residuals they can be regarded as observations of independent random variables and there is no further modeling to be done.

Neural Network Model

A Feed forward Neural network was fitted to the data with the help of SPSS 16.0 where values of the time series at 1^{st} , 2^{nd} and 3^{rd} lags were taken. The data was divided into 3 sets viz training, testing and holdout. 72.7 % observations were used for training, 25.9% for testing and 1.4% for forecasting

The information about the Neural network architecture are given in table -2 which shows that network has an input layer, a single hidden layer and an output layer. In the hidden layer there are 3 units and the activation function used is the hyperbolic tangent

The architecture of the network has been shown in the fig 7 light color lines show weights greater than zero and the dark color lines show weight less than zero

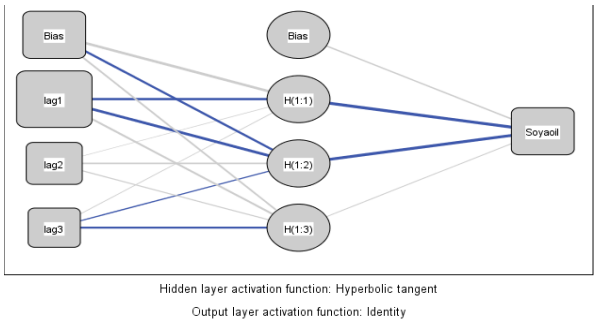

Fig 7: The architecture of the network fitted to time series of Soybean oil in Indore

The estimates of the weights and of the artificial neural network are shown in table:2.

Table- 2:The estimates of the weights and Bias of ANN fitted to Soybean oil in Indore

Predictor		Predicted			
		Hidden Layer 1			Output Layer
		H(1:1)	H(1:2)	H(1:3)	
Input Layer	(Bias)	.658	-.405	.345	
	lag1	-.635	-.676	.370	
	lag2	.040	.334	.166	
	lag3	.058	-.174	-.387	
Hidden Layer 1	(Bias)				.256
	H(1:1)				-1.310
	H(1:2)				-1.265
	H(1:3)				.102

CONCLUSION :

The ARIMA and ANN models were compared for their forecasting capabilities with the help of RMSE and MSE. The results are shown in table -3.

The one step ahead forecast for May 2010 (426.56) was best predicted by ANN model (426.01) and forecast by the ARIMA model (419.25).The two step ahead forecast for June 2010 (420.06) was best predicted by ARIMA followed by forecast by the ANN model (429.46).

Table 3- Observed and predicted prices of Soybean oil in Indore

Months	Observed (Prices Rs/ 10 lt)	Predicted(Prices Rs/ 10 lt)	
		ARIMA	ANN
May 2010	426.56	419.25	426.01
June 2010	420.06	418.75	429.46
July2010	438.36	421.97	422.41

MSE	323.78	343.07
RMSE	17.99	18.52
MAPE	1.92	2.00

The three step ahead forecast for July 2010 (438.36) was best predicted by ANN model (422.41) followed by forecast by the ARIMA model (421.97).It shows that therelativeperformance varies acrossforecasting horizons and different methodsperform best for different forecasting horizonsthis definitely point out the effect of time periodon the performance of the method.

REFERENCES :

1. Box, G.E.P. and Jenkins, G.M,. Time Series Analysis: Forecasting and Control.;revised ed Holden-Day, San Francisco. 1976
2. Gooijer De, Jan G. and Hyndman, Rob J.. 25 years of time series forecasting. Int. Journal of Forecasting. Elsevier. 2006;22(3): 443-473.
3. Kohzadi Nowrouz, Boyd Milton S.,Bahman Kermanshahi and Iebeling Kaastra,. A comparison of artificial neural network and time series models for forecasting commodity price. Neurocomputing, 1996;10: 169-18.
4. Tang, Z., Almeida De, C. and Fishcwick, P. A., Time series forecasting using neural networks vs. Box Jenkins methodology. Simulation., 1991;57(5): 303- 310.
5. Zoua , H.F., Xiaa, G.P. ,Yangc, F.T. and Wanga, H.Y., An investigation and comparison of artificial neural network and time series models for Chinese food grain price forecasting. Neurocomputing, 2007; **70**: 2913–2923.
6. McCulloch, W. S. and Pitts, W., A Logical Calculus of the Ideas Immanent in Nervous Activity. Bul. of Math.Biophysics.1943;5:115-133.
7. Rumelhart, D. E., Hinton, G. E., and Williams, R. J., Learning Internal Representations by Error Propagation, in Parallel Distributed Processing: Exploration in the Microstructure of Cognition. Cambridge, MA: MIT Press. 1986;1: 318-362.
8. Ripley, B. Neural Networks and Related Methods for Classification (with discussion). Journal of the Royal Statistical Society, 1994;56: 409-456.