Agris on-line Papers in Economics and Informatics

Volume V Number 3, 2013

A Study on Forecasting Prices of Groundnut Oil in Delhi by Arima Methodology and Artificial Neural Networks

G. C. Mishra, A. Singh

Department of Farm Engineering; Institute of Agricultural Sciences; BHU

Abstract

Forecasting of prices of commodities specially those of agricultural commodities is very difficult because they are not only governed by demand and supply but by so many other factors which are beyond control like weather vagaries, storage capacity, transportation etc. In this paper times series namely ARIMA (Autoregressive Integrated Moving Average) methodology given by Box and Jenkins has been used for forecasting prices of edible oils and this approach has been compared with ANN (Artificial Neural Network) methodology.

Key words

Forecasting, Prices, Groundnut oil, Delhi, ARIMA, ANN, Feed forward network.

Introduction

Price forecasting is very essential for planning and development and therefore it becomes pertinent to develop methods which helps the policy makers to have some idea about the prices of commodities in the future. One approach is to consider causes and their effects and the other approach is to forecast prices without taking in to consideration the causes. The time series approach to forecasting is one such approach which relies on the past pattern in a time series to forecast prices in the future. De Gooijer and Hyndman (2006) have provided an excellent review of time series methods in forecasting. There are many methods for analyzing a time series like exponential smoothing with a damped multiplicative trend Taylor (2003) etc., but one of the most simple and bench mark method is that of Box and Jenkins which is popularly known as ARIMA methodology. Dorfman and McIntosh (1990) suggest that structural econometrics may not give better results as compared to time series techniques even if the structural modelers are given the hard to find true model. The ARIMA approach has attracted researchers because it is a parsimonious approach which can represent both stationary and non-stationary stochastic processes as suggested by Harvey (1990). Numerous studies have shown that this univariate method is very effective as compared to some other multivariate methods like linear regression and vector autoregressive models. The problem with ARIMA methodology is that it assumes a linear structure of the process of which a particular times series is a realization, which is often not correct. 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. (1996). Apart from this artificial neural networks can also be used for classification problems as was shown by Ripley (1994). Artificial neural networks do not make any assumption about the process from which a particular time series has generated. Artificial neural networks effectively cover both linear and non linear processes. Combination of forecasts also increases the forecasting abilities of different methods as is being suggested by studies by Newbold et al. (1974), Zhang (2003), Zou et al. (2004), Hibon et al. (2005) and Makridakis and Hibon (2000) . In this paper time series of prices of groundnut oil in New Delhi from January 1994 to July 2010 has been analyzed with both the ARIMA methodology and artificial neural networks and the forecasting abilities of both the models has been compared.

Rest of the paper is organized as follows - in Section 2, the traditional univariate time series approach to forecasting is described and the neural network architecture that is designed for this study is discussed. It also discusses the evaluation methods for comparing the two forecasting approaches. Data and forecast procedure are discussed in Section 3. Section 4 shows the results obtained

from the ARIMA and the artificial neural network estimations. Section 5 shows conclusion.

Materials and Methods

Auto Regressive Integrated Moving Average (ARIMA) Time Series Model

Introduced by Box and Jenkins (1970), 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:

$$X_t = \theta 0 + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + e_t - \theta_1 e_{t-1} - \theta 2 e_{t-2} - ... - \theta e_{t-q} \text{ ... Eq} \quad (1)$$

Where X_t and e_t are the actual values and random error at time t, respectively, Φ_i (i = 1,2,……., p) and θ_j (j = 1,2,……, q) are model parameters, p and q are integers and often referred to as orders of autoregressive and moving average polynomials respectively. Random errors et are assumed to be independently and identically distributed with mean zero and the constant variance, σ_e^2. Similarly a seasonal model is represented by **ARIMA (p, d, q) x (P, D, Q)** model, where P = number of seasonal autoregressive (SAR) terms, D = number of seasonal differences, Q = number of seasonal moving average (SMA) terms. Basically this method has three phases: model identification, parameters estimation and diagnostic checking.

The ARIMA model is basically a data oriented approach that is adapted from the structure of the data itself.

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 (1943) for the first time proposed the idea of the artificial neural network 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. in 1986. 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. The ANN model performs a nonlinear functional mapping from the past observations ($X_{t-1}, X_{t-2}, ……., X_{t-p}$) to the future value X_t i. e.,

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

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 in which 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). However in practice, the choices of q and p are difficult.

To assess the prediction accuracy of the models understudy - the following Forecast Evaluation methods were applied:

Different criteria were used to make comparisons between the forecasting ability of the ARIMA time 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

$$AME = (1/T)\sum|P_t - A_t| \text{ ...Eq} \quad (3)$$

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 so 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

$$MSE = (1/T)\sum(P_t - A_t)^2 \text{ ... Eq} \quad (4)$$

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 4th criterion is RMSE which is the square root of MSE.

Data and Forecast Procedure

Monthly cash prices of groundnut oil in Delhi from April 1994 to July 2010 are used to test the prediction power of the two approaches. Data are obtained from the official website of ministry of agriculture. 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.

1. ARIMA Model

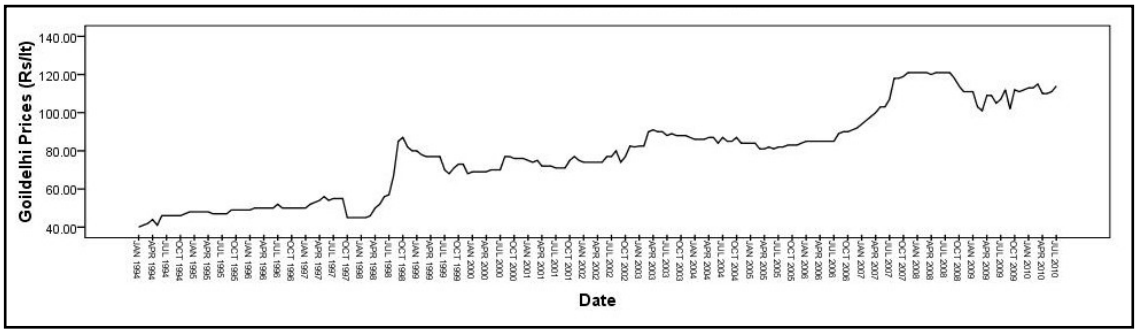
For fitting the ARIMA Model, the three stages of modeling as suggested by Box and Jenkins namely, identification, estimation and diagnostic checking was undertaken. Identification was done after examining the autocorrelation function and the partial autocorrelation function. After that, estimation of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

In Figure 1 shows the time plot of prices of the groundnut oil in Delhi. By looking at 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, autocorrelation function should also be seen. Box and Jenkins suggested that the most autocorrelations which may safely be examined is about one-fourth of the number of observations. So in the present case 50 autocorrelations were calculated.

In figure 2 is shown the autocorrelation function of the time series and it certainly shows that the series is not stationary because autocorrelation coefficients does 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.

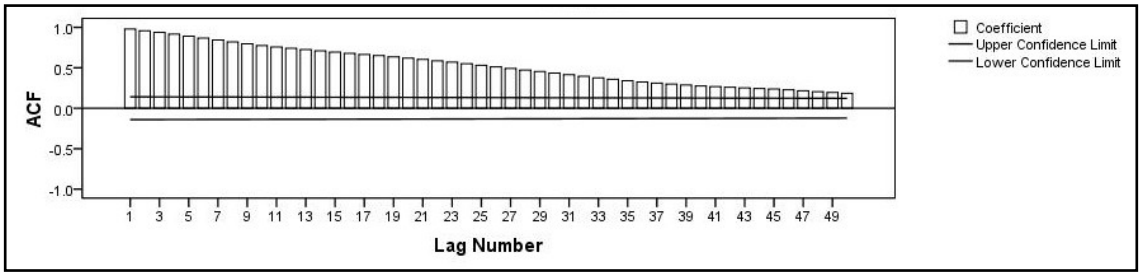
Figure 3 shows the time plot of the differenced series and it clearly depicts that the series has now become mean stationary. By looking at the variance of the series log transformation of the data was taken. The observations seem to fluctuate around a fixed mean, and the variance seems to be varying over time. However, the judgment about stationarity of the mean was withheld until the estimated ACF and perhaps some estimated AR coefficients were examined.

In figure 4 autocorrelation function and partial autocorrelation function (PACF) of the differenced series are shown. The autocorrelations decay to



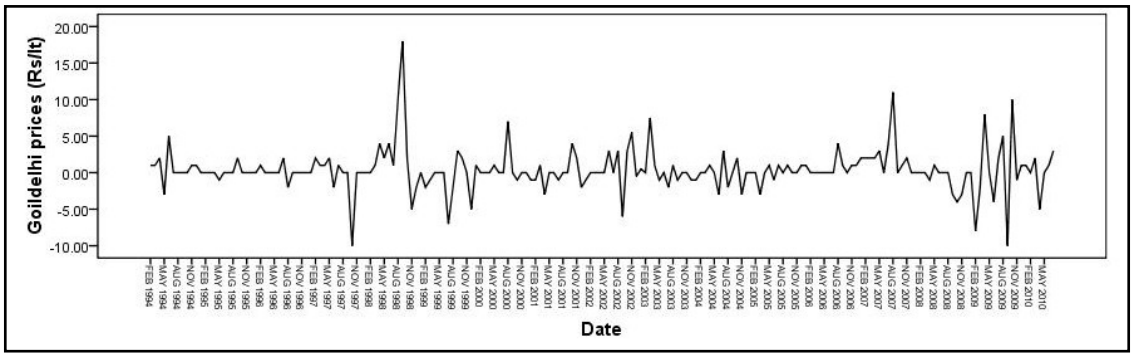
Source: Processing with use Statistical Package for Social Sciences

Figure 1: Time plot of the prices of groundnut oil in Delhi.



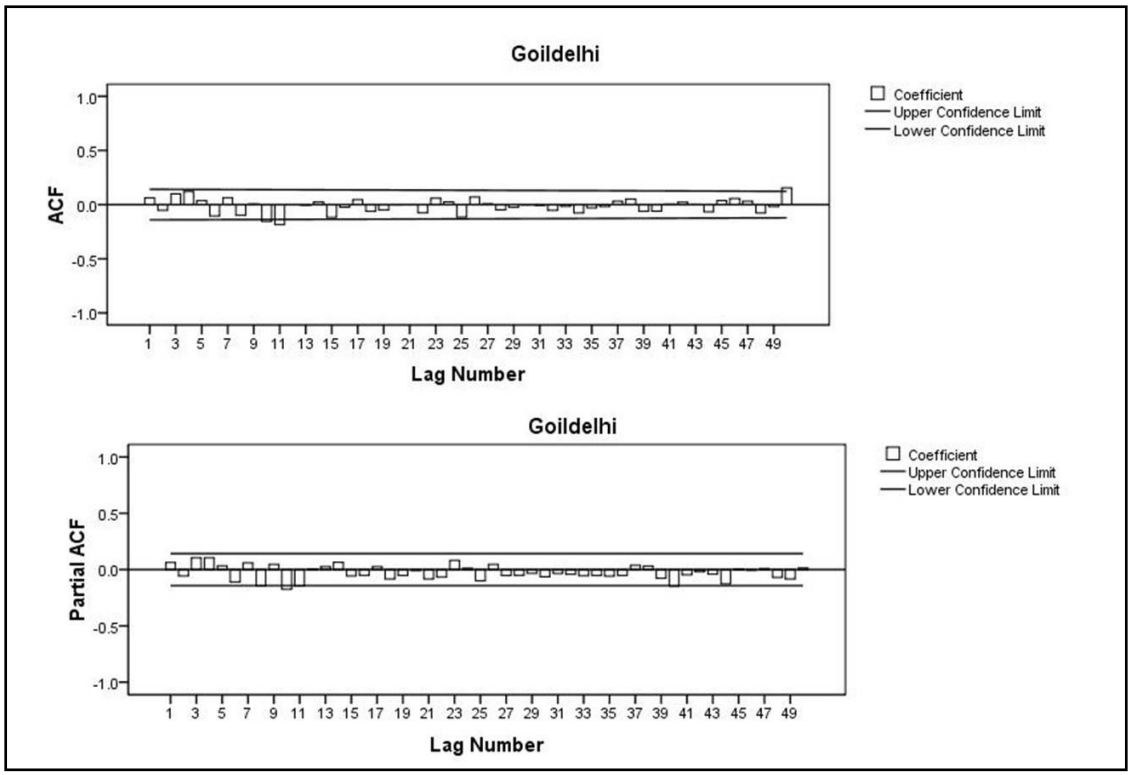
Source: Processing with use Statistical Package for Social Sciences

Figure 2: Autocorrelations at different lags



Source: Processing with use Statistical Package for Social Sciences

Figure 3: Transforms: difference (1).



Source: Processing with use Statistical Package for Social Sciences

Figure 4: ACF and PCF of the differenced series.

statistical insignificance rather quickly. It was concluded that the mean of the series is probably stationary. The data series and the autocorrelations didn't indicate to the presence of seasonality. However spectral density of the price by frequency was observed and there was no seasonality in the data. The PACF are significant at around lag 10 and 11.

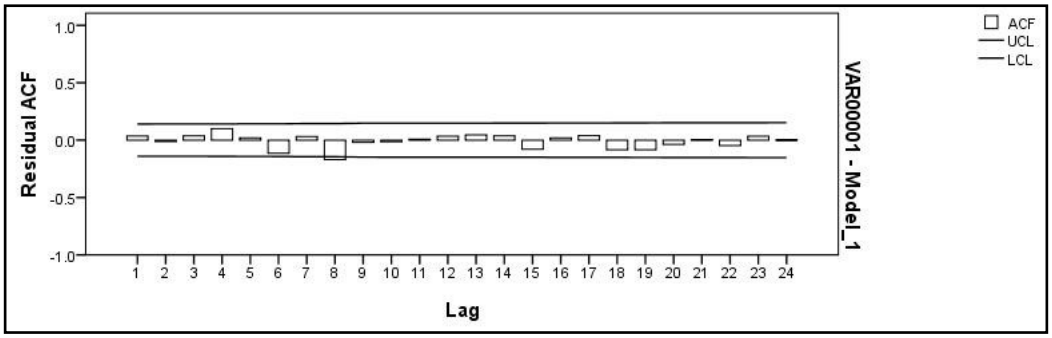
Once the time series has become stationary Using Expert Modeler option in SPSS, the ARIMA model was estimated. After going through these stages ARIMA (0,1,11) model was found to be the best among the family of ARIMA models. ARIMA Model parameters and model Fit statistics are given in the Table 1.

In the Table 1, it is shown that constant = 0.005 with a S.E. of .001 which was significant at 1% level of significance. Although ARIMA (0, 1, 11) was found to be the best model only moving average (MA) terms at lag 10 and lag 11 were found to be statistically significant at 1% level of significance and therefore only significant values are being shown in the Table 1., with an estimate of 0.195 at lag 10 and 0.37 at lag 11 and a standard error of 0.072 and 0.073 at lag 10 and lag 11 respectively .

	Estimate	SE	t	Sig	Model Fit Statistics	
Constant	0.005	0.001	4.282	0	Stationary R Squared	0.139
Difference	1				R Squared	0.985
MA Lag 10	0.195	0.072	2.697	0.008	RMSE	2.821
Lag 11	0.37	0.073	5.08	0	MAPE	2.227
					MAE	1.736
					Normalized BIC	2.155

Source: Processing with use Statistical Package for Social Sciences

Table 1.



Source: Processing with use Statistical Package for Social Sciences

Figure 5: ACF of the residuals.

This model satisfies the stationarity requirement $\theta_{11} + \theta_{10} < 1.0$. Also θ10 , and θ_{11} are highly significantly different from zero since its absolute t-value of 2.697 at lag 10 and absolute t-value of 5.08 at lag 11 which is greater than 2.0. Also R^2 value is 0.985 and RMSE, MAPE, MAE, BIC are 2.821, 2.227, 1.736 and 2.155 respectively showing satisfactory model fitting.

At the diagnostic checking stage residual were examined and their autocorrelation coefficients were found to be non significant (Figure 5). Which shows that the model is satisfactory.

To determine if model is statistically adequate, the random shocks for 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 were independent. With 150 residuals about 24 residual autocorrelations were examined. The residual ACF appears below the estimation results in the Figure 5. 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. If there is no dependence among the residuals then we can regard them as observations of independent random variables and there is no further modeling to be done.

Since, θ_{10} and θ_{11} meets the stationarity requirement and is statistically different from zero, constant is significant and the shocks appear to be independent according to the t-tests. Thus, forecasting can be done.

2. 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 1st, 2nd and 3rd lags were taken for forecasting. The data was divided into 3 sets viz. training, testing and holdout. 81.6 % observations were used for training, 16.8% for testing and 1.5% for forecasting (Table 2).

		N	Percent
Sample	Training	160	81.60 %
	Testing	33	16.80 %
	Holdout	3	1.50 %
	Valid	196	100.00 %
	Excluded	0	
	Total	196	

Source: Source: Processing with use Statistical Package for Social Sciences

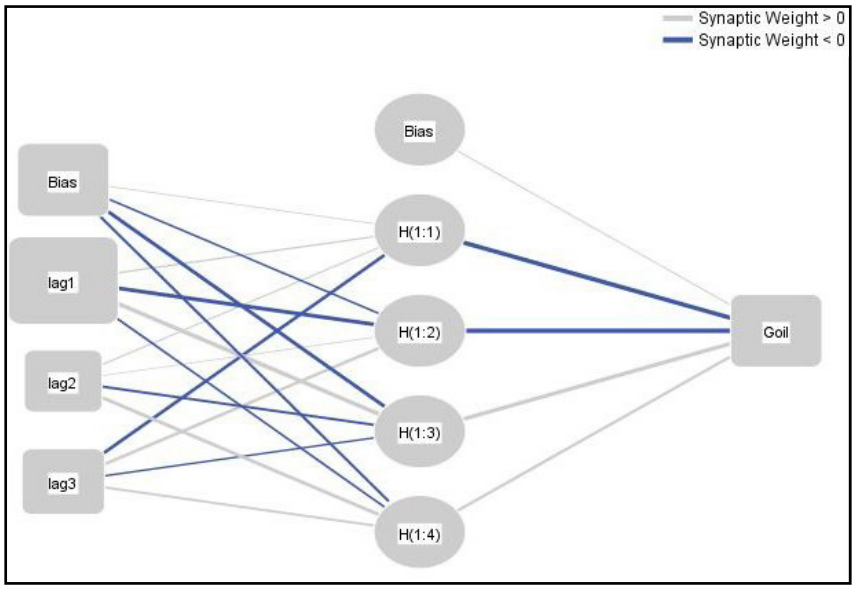
Table 2: ANN Case processing summary of groundnut oil in Delhi.

The information about the neural network

Input layer	Covariates	Lag1, lag2, lag3
	No. of units	3
	Rescaling methods of covariates	Standardized
Hidden Layers	No. Of hidden layers	1
	No. of units in hidden layers	4
	Activation Function	Hyperbolic tangent
Output Layer	Dependent variables	1
	Number of units	1
	Rescaling methods for scale dependents	Standardized
	Activation function	Identity
	Error function	Sum of squares

Source: Source: Processing with use Statistical Package for Social Sciences

Table 3: Network information for groundnut oil in Delhi



Source: Processing with use Statistical Package for Social Sciences

Figure 6: Hidden layer activation function: Hyperbolic tangent.
Output layer activation function: Identity.

Training	Sum of Squares Error		3.795
	Relative Error		0.048
	Stopping Rule Used	Maximum number of epochs (100000)	exceeded
Testing	Sum of Squares Error		1.021
	Relative Error		0.461
Holdout	Relative Error		0.454

Source: Processing with use Statistical Package for Social Sciences

Table 4: The training summary and the fit statistics of ANN of groundnut oil in Delhi.

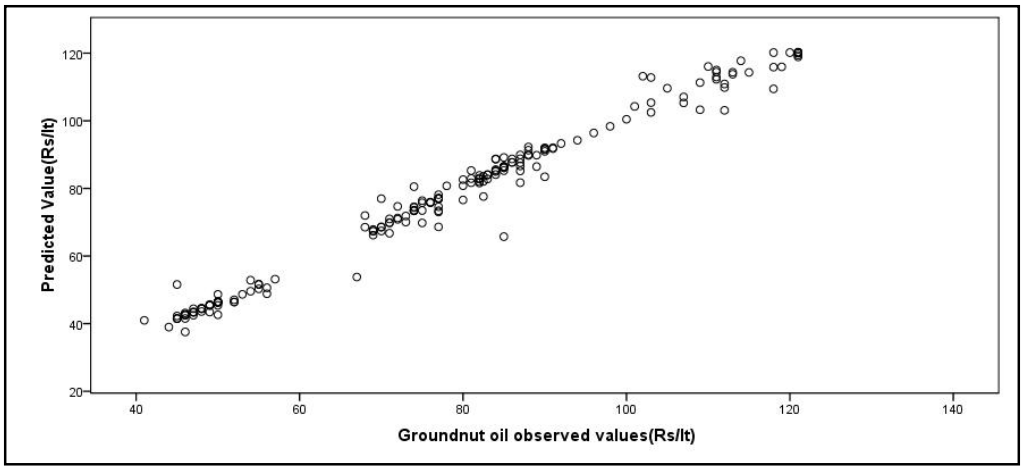
architecture is given in Table 3 which shows that network has an input layer, a single hidden layer and an output layer. In the hidden layer there are 4 units and the activation function used is the hyperbolic tangent

The architecture of the network has been shown in the Figure 6, light color lines show weights greater than zero and the dark color lines show

		Predicted				
		Hidden Layer 1				Output Layer
Predictor		H(1:1)	H(1:2)	H(1:3)	H(1:4)	Goil
Input Layer	(Bias)	0.021	-0.197	-0.562	-0.272	
	lag1	0.086	-0.646	0.712	-0.235	
	lag2	0.073	0.018	-0.253	0.468	
	lag3	-0.429	0.363	-0.102	0.237	
Hidden Layer 1	(Bias)					0.024
	H(1:1)					-1.548
	H(1:2)					-1.704
	H(1:3)					0.66
	H(1:4)					0.349

Source: Processing with use Statistical Package for Social Sciences

Table 5: The estimates of the weights and Bias of ANN fitted to groundnut oil in Delhi.



Source: Processing with use Statistical Package for Social Sciences

Figure 7: Observed vs predicted prices

weight less than zero.

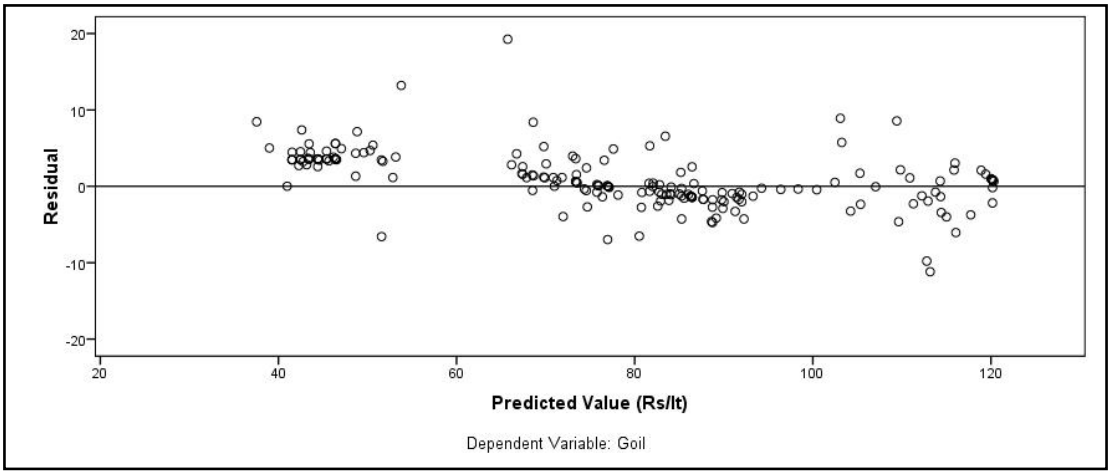
The training summary and the fit statistics for the training, testing and the holdout sets are given in Table 4.

The estimates of the weights and bias are given in Table 5. This table shows the value of weights from input to the hidden layer and from the hidden layer to the output layer. H(1:1) means Hidden layer 1 and 1St neuron. The weight attached to the neuron from bias is 0.021, from lag 1 is 0.086, from lag 2 is 0.073 and from lag 3 is -0.429. H (1:2) means Hidden layer 1 and 2nd neuron. The weight attached to the neuron from bias is -.197, from lag 1 is -.646 from lag 2 is.018 and from lag 3 is .363. H (1:3) means hidden layer 1 and 3rd neuron. The weight attached to the neuron from bias is -.562, from lag 1 is .712 from lag 2 is-.253 and from lag 3 is -.102. H (1:4) means Hidden layer 1 and 4th neuron. The weight attached to the neuron from bias is -.272, from lag 1 is -.235 from lag 2 is.468 and from lag 3 is .237.

The weights from the hidden layer to the output layer for bias .024 and from 1st neuron in the hidden layer to the output is -1.548, from 2nd neuron in the hidden layer to the output is -1.704. from 3rd neuron in the hidden layer to the output is .660 and from 4th neuron in the hidden layer to the output is .349.

The observed vs. the predicted graph is shown in the Figure 7 which depicts that except for few outliers it is a straight line. It indicates almost one to one correspondence among the observed and predicted values. Hence it can be inferred that the performance of ANN is satisfactory.

The residual vs. predicted graph (Figure 8) also



Source: Processing with use Statistical Package for Social Sciences

Figure 8: Residuals vs predicted plot

Months	Observed(Prices Rs/lt)	Predicted(Prices Rs/lt)				
		ARIMA	ANN	Combined		
				Equal weights	Weights = 1/RMSE	Weights = 1/MAPE
V.10	110	111.89	111.21	111.55	111.54	111.59
VI.10	111	111.75	112.08	111.915	111.92	111.89
VII.10	114	114.01	112.86	113.48	113.47	113.55
MSE		4.1446	3.93	3.5	3.51	3.53
RMSE		2.036	1.98	1.86	1.87	1.88
MAPE		0.83	1.02	0.89	0.89	0.88

Source: Processing with use Statistical Package for Social Sciences

Table 6: Observed and predikted prices of Groundnut oil in Delhi.

shows that the residual do not follow a definite pattern and therefore are not correlated. If there is no dependence among the residuals then we can regard them as observations of independent random variables and believe that the ANN is satisfactory.

Results and discussion

The ARIMA and ANN models were compared for their forecasting capabilities with the help of RMSE and MSE. The results are shown in the Table 6.

The one step ahead forecast for May 2010 (110) was best predicted by ANN model (111.21) followed by combined forecast with weights equal to 1/ RMSE (111.54), followed by combined forecast with equal weights (111.55), by combined forecast with weights equal to 1/MAPE (111.59) and forecast by the ARIMA model (111.89).

The two step ahead forecast for June 2010 (111) was best predicted by ARIMA model (111.75) followed by combined forecast with weights equal to 1/MAPE (111.89) (111.54), by combined forecast with equal weights(111.915) , by combined forecast with weights equal to 1/RMSE (111.92) and forecast by the ANN model (112.08).

The three step ahead forecast for July 2010 (114) was best by ARIMA model (114.01) followed by combined forecast with weights equal to 1/MAPE (113.55), by combined forecast with equal weights(113.48), by combined forecast with weights equal to 1/RMSE (113.47) and forecast by the ANN model (112.86).

Overall the forecast by ARIMA model was found to be the best with MAPE(0.83),RMSE (2.036), MSE (4.1446) followed by combined forecast with weights equal to 1/MAPE with MAPE(0.88),

RMSE(1.88), MSE(3.53), by combined forecast with equal weights with MAPE(0.89), RMSE (1.86), MSE(3.50), by combined forecast with weights equal to 1/RMSE with MAPE(0.89), RMSE(1.87), MSE(3.51) and forecast by the ANN model with MAPE(1.02), RMSE (1.98), MSE(3.93).

Conclusion

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.

Further, forecasting of prices can be of great help to poor farmers in deciding what to cultivate and when to sell. This will certainly help in reducing the exploitation of farmers by the middlemen and will uplift the socio-economic status of the poor farmers.

This study compared neural network and ARIMA models to forecast monthly prices of groundnut oil in Delhi one of the major Indian markets. It is well known that forecasting of prices of agricultural commodities is always and will remain difficult because such data are greatly influenced by economical, political, international and even natural shocks. Neural networks have the ability to model nonlinear patterns and learn from the historical data. ARIMA models were used as a benchmark. In the literature of time series forecasting with neural networks, most studies use the ARIMA models as the benchmark to test the effectiveness of the ANN model like Zoua et al. (2007) and Tang et al (1991). Monthly data was used from 1994 to 2010. The mean squared error, root mean square error and mean absolute percent error were all lower on average for the ARIMA forecast than for the neural network. Following conclusions were drawn from the study.

- Accuracy depends upon the forecasting horizon-The relative performance varies across forecasting horizons and different methods perform best for different forecasting horizons this definitely point out the effect of time period on the performance of the method. This can be seen from the fact that for May 2010, forecast by ANN model was found to be better but for two and three step ahead forecasts i.e. for June 2010 and July 2010, ARIMA model performed better than the ANN.
- Performance ranking varies by metric. The rankings of the contestants based upon the MAPE, MSE, and RMSE each result in different relative performances of the methods used across all datasets and data conditions. This can be inferred from the fact that for the overall performance we compare the methods by looking at the values of RMSE than ANN model performed better but if we check the value of MAPE, the ARIMA model performed better. However, some methods performed consistently well on multiple metrics, and vice versa, increasing the confidence in their relative performances and predictive capabilities.

Corresponding author:
Dr. Abhishek Singh, Assist. Professor
Department of Farm Engineering, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi,Uttar Pradesh, India,221005

Phone: 9451526775, E-mail: asbhu2006@gmail.com

References

[1] Box, G. E. P., Jenkins, G. M. Time Series Analysis: Forecasting and Control. revised ed Holden-Day, San Francisco. 1976. ISBN:0816210942.

[2] De Gooijer, Jan. G., Hyndman, R. J. 25 years of time series forecasting. International Journal of Forecasting.Elsevier. 2006, vol. 22(3), pages 443-473. ISSN:0169-2070.

[3] Dorfman, J. H., Mcintosh, C.S. Results of a price forecasting competition, American J. Agricultural Economics. 1990,72, 804-808. ISSN:0002-9092.

[4] Harvey, A.C. The Econometric Analysis of Time Series.1990 (MIT Press,). ISBN-0860031926

[5] Hibon, M., Evgeniou, T. To combine or not to combine: selecting among forecasts and their combinations. International Journal of Forecasting. 2005, 21, 15–24. ISSN:0169-2070.

[6] Kohzadi N., Boyd M. S., Bahman K., Iebeling K. A comparison of artificial neural network and time series models for forecasting commodity price. Neurocomputing. 1996,10., 169-181. ISSN:0925-2312.

[7] Makridakis, S., Hibon M. The M3-Competition: results, conclusions and implication., International Journal of Forecasting. 2000, 16, 451–476. ISSN:0169-2070.

[8] Mcculloch, W. S., Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics.1943, 5, 115-133. ISSN:0007-4985.

[9] Newbold, P., Granger C. W. J. Experience with forecasting univariate time series and the combination of forecasts. Journal of the Royal Statistical Society (A). 1974, 137, 131–165. ISSN:0035-9238

[10] Ripley, B. Neural Networks and Related Methods for Classification. (with discussion). Journal of the Royal Statistical Society. 1994, Ser. B, 56, 409-456. ISSN:1467-9868.

[11] Rumelhart, D. E., Hinton, G. E., Williams, R. J. Learning Internal Representations by Error Propagation, in Parallel Distributed Processing: Exploration in the Microstructure of Cognition., Cambridge,MA: MIT Press. 1986. Vol.l. pp. 318-362. ISBN:0-262-68053-X.

[12] Tang, Z., De Almeida, C., Fishcwick, P. A. Time series forecasting using neural networks vs. Box Jenkins methodology. Simulation. 1991,57 5, p. 303-310. ISSN:0037-3497.

[13] TAYLOR, J.W. Exponential smoothing with a damped multiplicative trend, International Journal of Forecasting. 2003,19, 273–289. ISSN:0169-2070.

[14] Zhang G. P. Times series forecasting using a hybrid ARIMA and neural network model. Neurocomputing.; 2003,50,159–75. ISSN:0925-2312.

[15] Zou, H., Yang, Y. Combining time series models for forecasting. International Journal of Forecasting. 2004, 20, 69–84. ISSN:0169-2070.

[16] Zoua, H. F. , Xiaa, G. P. , Yangc, F. T., 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. ISSN:0925-2312.