Journal of Academia and Industrial Research (JAIR)
Volume 3, Issue 1 June 2014

ISSN: 2278-5213

RESEARCH ARTICLE

Soybean Price Forecasting in Indian Commodity Market: An Econometric Model

Rajesh Panda
Symbiosis Institute of Business Management, Bengaluru (A constituent of Symbiosis International University),
95/1, 95/2, Electronic City Phase-1, Hosur Road, Bengaluru-560100, Karnataka, India
rajeshpanda.80@gmail.com, director@sibm.edu.in

Abstract

Soybean is one of the important oil seeds cultivated in India and has multiple usages as whole soybean, soybean meal, soybean protein products and soy oil. However, soybean is usually harvested once year leading to an uneven supply across months and price fluctuation throughout the year. With evolution of the commodity stock exchanges in India, it creates avenues for speculations for the future price of the soybean in the absence of a proper forecasting system. This leads to price asymmetry, speculators gaining at the cost of traders and farmers. This study tries to identify a forecasting model that best replicates the actual situation so as to minimize the speculative gains. Time series method of forecasting has been used for building the forecasting model. ARIMA (1,1,0) and additive seasonality best represented the past data. The same model was extrapolated and compared with actual data to justify the fitness of the model.

Keywords: Soybean, forecasting, ARIMA, seasonality, speculative gains, stock exchanges, price asymmetry.

Introduction

Soybean is one of the most important oilseeds in India. The cost of cultivation of soybean is cheaper compared to other oil seeds. Soybean is adaptable to different types of climatic conditions and soil structures. Despite such benefits, soybean became popular as an oil seed only in 18th century though it has been cultivated for more than 4800 years. India is the 5th largest producer of Soybean after USA, Brazil, China and Argentina (USDA Economic and statistics System, 2000). The area under cultivation of soybean has increased phenomenally in India in last twenty years. In India, soybean seeds are sown in the month of June and harvested in December. However, soybean demand is almost even throughout the year because of its versatile usage. Continuous demand throughout the year coupled with uneven supply leads to huge variation in soybean prices. Moreover, with the evolution of commodity exchanges, this has given rise to speculation for the price of soybean seeds. Under such circumstances, absence of any proper price forecasting mechanism may harm both the harvesters and the traders. The benefit of this price asymmetry may go to a new breed of speculators at the cost of the harvesters and traders. This necessitates development of a price forecasting model for soybean seeds to reduce the uncertainty and help both farmers and traders. The increased importance of soybean comes from its versatile end uses. Soybean uses can be as whole soybean, soybean meal, soybean protein products and soy oil (Soy Protein Council, 1987). In addition to its use as human food, soybean can also be used for fodder and for different industrial usage. Whole soybeans can be cooked and used in sauces, stews and soups, soy flour is used in various products in baking and roasted soybean is used in different confectioneries. Soybean meal is the by-product of soya oil production. Soybean meal is used for preparation of fodder and fertilizer. Soybean meal is used in premium fodder because of its high digestibility and high calorie content. Soybean protein products include soy flour, concentrates and isolates. Soybean seed prices follow huge monthly variation in the spot market. December being the harvest season for soybean, the supply in the market is high and hence, the prices is generally low from December to March and the prices usually peak in the months of September and October. Unlike a stock, which represents equity in a company and can be held for a long time, if not indefinitely, futures contracts have finite lives. They are primarily used for hedging commodity price-fluctuation risks or for taking advantage of price movements, rather than for the buying or selling of the actual cash commodity. The word "contract" is used because a futures contract requires delivery of the commodity in a stated month in the future unless the contract is liquidated before it expires. To work in futures market and gain from it, accurate forecasting is very necessary. For an accurate forecasting, a good understanding of the factors influencing the price of soybean like harvest season, demand for the final product, demand of substitutes etc. are essential. Unavailability of the relevant data and forecasting measures hinders traders and hedgers to enter into the futures market in India. This leads to biased speculation which leads to unavoidable fluctuations in price. This reduces the predictability of the market and makes it less useful for the genuine hedgers for trading and also leads to inappropriate price realization for farmers. The asymmetry in information and the highly fluctuating

prices creates problems both for the harvesters and the traders. Though people still trade in the futures market most of them are still not aware of the underlying basis risk and the methods of reducing it through the futures trading. This is because of the lack of predictability in the market due to which the basis risk also becomes unpredictable. This study provides an analysis of market prices of the soybean seed and attempts to bring out the underlying patterns in the price data for near about accurate forecasting with the following objectives:

- To understand the problems and uncertainties in forecasting prices of soybean in India.
- To understand the various factors influencing the fluctuation of the soybean prices, and the effect of their trend and seasonality components on the past prices.
- To evolve a model for most possible accurate forecasting of soybean seed prices.

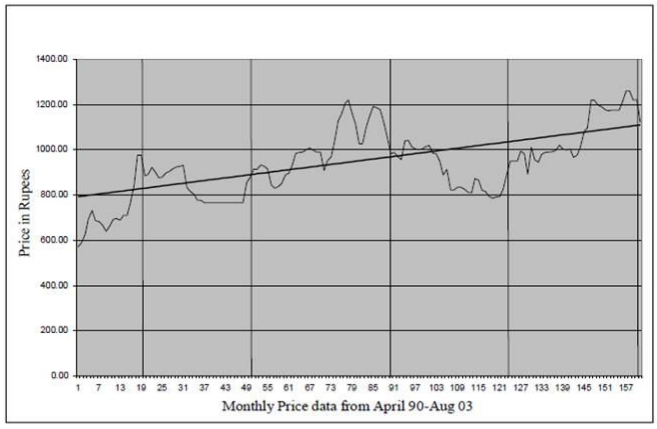
Materials and methods

Assessment of forecasting needs: There was a need to assess the necessity of forecasting of soybean prices to the harvesters, hedgers, traders and the commodity exchanges. Discussions with commodity traders revealed that neither the variation in the price nor the effect of the factors influencing the price is important for them. They need the absolute forecasted prices, the availability of which would ensure a price reflecting the market reality and reduce price distortions and chances of speculation. This will help the traders as well as the producers in hedging their risk and a commodity exchange will be able to restrict the abnormal profits gained by the speculators. This will make the trading predictable and would ensure that no supernormal profits would be gained by a one or a group of speculators. This would help more buyers and sellers to trade in commodity and ultimately benefiting the farmer by reducing the information asymmetry.

Data collection: The monthly average spot price data of soybean for 13 years (from April 90 to August 2003) collected from CMIE has been used of forecasting and compared with the actual spot price for subsequent eight years. Time series data analysis has been used to forecast the monthly soybean seed prices. The data on market situation, futures trading was collected from electronic sources and other related sites of the soybean market.

Time series analysis: Time series analysis has been chosen to analyze the data because the number of data points is about 160 in which case time series analysis would produce a better and required absolute values of forecast. Causal model analyses the influence of individual factors on the price. This model has not been used for analysis because market conditions and other factors influencing soybean prices change frequently and hence, the prediction of probable factors would be difficult.

Fig. 1. Soybean price.



Source: CMIE database.

Moreover, the forecast model would become obsolete after a short span of time if a new factor is found to influence the price or if one of the factors stopped influencing the price in the model. Adding to this, the availability of the data on these factors would be difficult. Hence, time series analysis was preferred over causal model. The data (soybean price) was plotted against time (months) to observe trend and seasonality (Fig. 1).

Data analysis: Soybean is an annual crop and follows a specific pattern in price where, the price of the seed falls during the harvesting season (November to February) and increases during the lean months' period (July to September). The inter year variations occur due to changes in weather, Government policies, harvest of competitive crops, etc. The pattern in inter year variations can be identified by the trend and cycle while the intra year variation can be computed in the form of normalized seasonality indices. The process of analysis of the data and built up of the model is depicted in Fig. 2.

Fig. 2. Flow chart for data analysis
(Proposed model by the author).

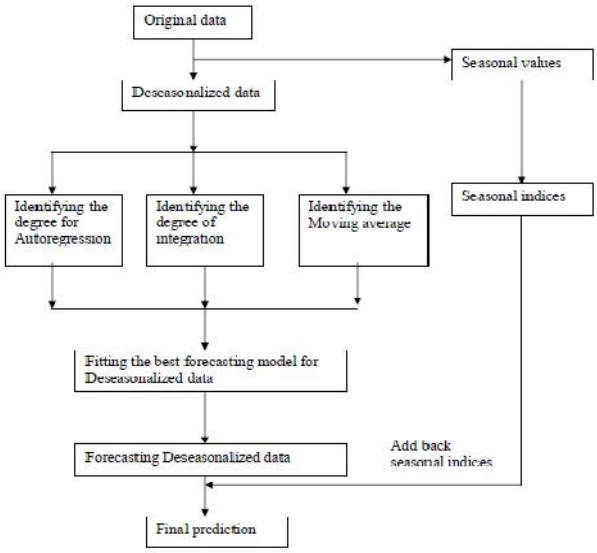


Table 1. Deseasonalization.

		Price	Uncentered T+C=12MA	Centered MA T+C	Seasonality + Error
Apr-90	1	571.89			
May-90	2	588.71			
Jun-90	3	625.41			
Jul-90	4	694.22			
Aug-90	5	731.68			
Sep-90	6	687.34	661.41		
Oct-90	7	681.99	671.09	666.25	15.74
Nov-90	8	667.46	681.16	676.13	-8.67
Dec-90	9	639.94	688.23	684.70	-44.76
Jan-91	10	661.34	694.35	691.29	-29.95
Feb-91	11	691.16	703.84	699.09	-7.93
Mar-91	12	695.75	727.86	715.85	-20.10
Apr-91	13	688.10	752.39	740.13	-52.02
May-91	14	709.51	770.55	761.47	-51.96
Jun-91	15	710.28	791.51	781.03	-70.75
Jul-91	16	767.62	813.24	802.37	-34.76
Aug-91	17	845.60	830.89	822.06	23.54
Sep-91	18	975.58	845.92	838.40	137.17
Oct-91	19	976.34	861.85	853.89	122.46
Nov-91	20	885.36	877.52	869.69	15.67
Dec-91	21	891.48	893.64	885.58	5.89
Jan-92	22	922.06	906.07	899.86	22.20
Feb-92	23	902.95	912.63	909.35	-6.40
Mar-92	24	876.19	908.55	910.59	-34.41
Apr-92	25	879.24	904.79	906.67	-27.43
May-92	26	897.59	900.40	902.59	-5.00
Jun-92	27	903.71	894.15	897.27	6.44
Jul-92	28	916.71	884.34	889.25	27.46
Aug-92	29	924.35	873.96	879.15	45.20
Sep-92	30	926.65	865.67	869.81	56.83
Oct-92	31	931.23	856.12	860.89	70.34
Nov-92	32	832.61	845.03	850.57	-17.97
Dec-92	33	816.55	833.43	839.23	-22.68
Jan-93	34	804.32	820.76	827.09	-22.78
Feb-93	35	778.32	807.44	814.10	-35.78

Values similarly calculated till Aug 2003.

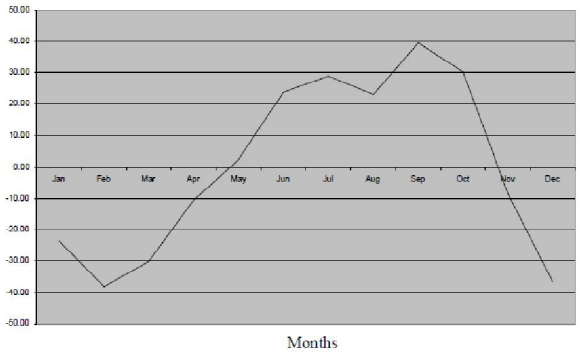
Results and discussion

The proposed model can be described in following nine steps:

Step 1: Deseasonalizing data: The data was deseasonalized by removing the seasonality from the data. The trend and cycle were computed using 12 months centered moving average (12 CMA) method since soybean is an annual crop (12 months). The trend and cycle part of the data were removed, leaving the seasonal and the irregular or error terms. The numerical method of deseasonalizing has been shown in Table 1.

Step 2: Computation of seasonal indices: It is assumed that the seasonal component is constant year on year. So we need to calculate seasonal indices. All seasonal values for a given month were gathered and the average was computed, e.g. the seasonal index of January is the average of all the values for Jan. Seasonal component is constructed by stringing together the seasonal indices for each year of data as shown in Fig. 3. The calculation of the seasonal indices is shown in Table 2.

Fig. 3. Seasonality indices.



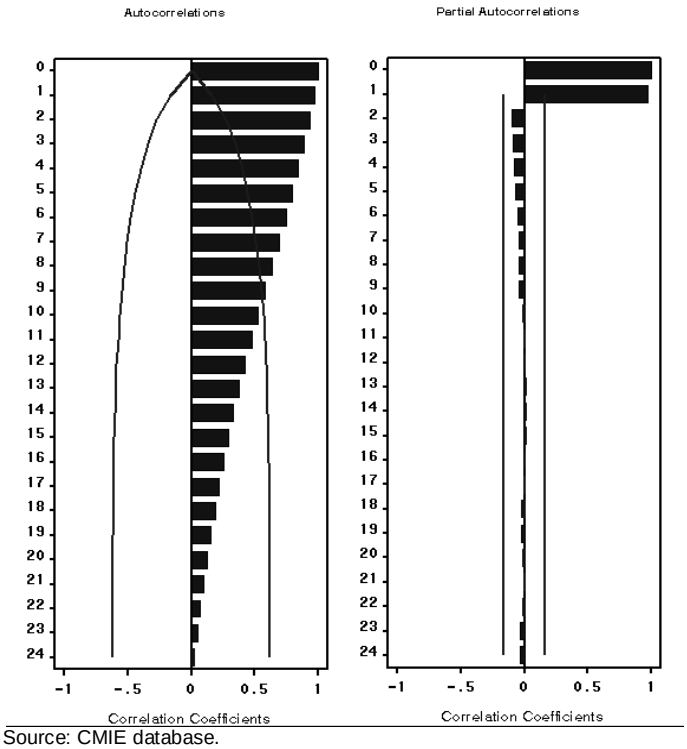
Source: CMIE database.

Step 3: Identification of the degree of autocorrelation: It can be seen that in the Auto Correlation Function (ACF) graph that the correlation is gradually drooping down with the increase in the number of lags while the Partial Auto Correlation Function (PACF) graph cuts off and shows that the value of first lag is significant (Fig. 4). So, it can be inferred that the series has an autocorrelation of first degree.

Table 2. Seasonal indices.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
										15.74	-8.67	-44.76
	-29.95	-7.93	-20.10	-52.02	-51.96	-70.75	-34.76	23.54	137.17	122.46	15.67	5.89
	22.20	-6.40	-34.41	-27.43	-5.00	6.44	27.46	45.20	56.83	70.34	-17.97	-22.68
	-22.78	-35.78	-23.89	-22.43	-12.65	-7.65	-3.82	-1.59	-0.51	-3.76	-12.23	-23.22
	-35.68	-48.84	-62.60	14.59	28.00	59.48	51.45	63.90	50.56	30.74	-44.31	-63.17
	-61.52	-52.53	-22.04	-21.34	5.00	43.07	35.36	27.40	33.83	35.01	18.57	-0.06
	-13.06	-110.86	-85.44	-84.93	-30.45	56.39	83.18	121.88	121.85	43.99	-16.06	-113.95
	-115.54	-34.79	28.67	90.89	98.79	97.90	50.65	-16.31	-84.04	-64.45	-67.19	-66.74
	27.75	38.67	9.62	0.57	-0.54	5.51	25.04	43.29	27.56	39.47	19.02	-27.08
	13.99	-59.89	-46.54	-24.75	-13.86	-14.72	-20.48	-15.55	49.92	45.20	3.79	-6.02
	-34.28	-54.98	-57.09	-64.10	-46.32	9.53	57.66	41.48	27.97	58.14	32.02	-70.02
	44.03	-14.81	-31.22	1.40	7.20	4.08	3.54	9.94	28.48	4.71	-7.80	-16.60
	-73.88	-81.71	-61.26	-7.68	-5.42	103.28	86.65	47.47	25.45	-4.17	-23.03	-27.56
	-27.68	-25.39	15.75	61.15	58.71	16.59	13.52	-91.74				
Average	-23.57	-38.10	-30.04	-10.47	2.42	23.78	28.88	22.99	39.59	30.26	-8.32	-36.61
Normalized	**-23.64**	**-38.16**	**-30.11**	**-10.54**	**2.36**	**23.71**	**28.81**	**22.93**	**39.52**	**30.20**	**-8.39**	**-36.68**

Fig. 4. The ACF and PACF of the deseasonalized data.



Source: CMIE database.

Step 4: Identification of moving average: It can be seen from the ACF graph that the correlation is gradually drooping down and do not have significant spikes through a definite number of lags while the PACF cuts off. So it can be inferred that it does not follow a moving average process.

Step 5: Identification of degree of integration: The degree of integration can be detected by finding out the number of differences required to make the data stationary. For testing the stationarity of the deseasonalized data, Dickey Fuller test was done (Gujarati and Sangeetha, 2007). The test was conducted under different forms as mentioned below.

(i) Y_t is a random walk

Equation: $\Delta Y_\tau = \delta Y_{\tau-1} + u_\tau$ (1)

Null hypothesis: $\delta = 0$: Y_τ is non-stationary.

Alternate hypothesis: δ is negative: Y_t is stationary.

ΔY_t was regressed upon Y_{t-1} and the calculated t value was found to be 4.016027. This model has to be ruled out because the coefficient of Y_{t-1} has a value of 0.003611, which is positive. But since $\delta = (\rho - 1)$, a positive δ would imply that $\rho > 1$ which is theoretically impossible. The regression output is shown in Table 3.

(ii) Y_t is a random walk with drift

Equation : $\Delta Y_t = \beta_1 + \delta Y_{t-1} + u_t$ (2)

In this regression, the coefficient of Y_{t-1}, that is δ = -0.00443, which is negative implying that the value of ρ is below 1. The t value so obtained can be compared only with the Dickey Fuller (DF) critical values (τ statistics) as the differenced data does not follow 't' statistics. In this case, the t value is within the critical DF value of -2.89 at 5% level of significance. Hence, the null hypothesis is accepted i.e. $\delta = 0$ reflecting that the data is non stationary. The regression output is shown in Table 4. To make the data stationary, it was differenced on the first order and was tested for stationarity again as follows.

Y_t is a random walk

Now stationarity test for the first order differenced data is calculated as follows:

$\Delta Y_\tau - \Delta Y_{\tau-1} = \delta Y_{\tau-2} + u_\tau$ (3)

In this regression, the coefficient of Y_{t-2}, that is δ = -0.0476, which is negative implying that the value of ρ is below 1. In this case, the t value is -1.96511, which is negative and the more than the critical DF table values of -1.95 at 5% significance level (in absolute terms). Hence, we reject the null hypothesis ($\delta = 0$: Y_t is non-stationary), so it can be concluded that the data is stationary. It can be inferred that the data follows an integration of first order. The regression output is shown in Table 5.

Table 3. Random walk model without drift.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
X Variable 1	**0.003611**	0.000899	**4.016027**	**9.24889E-05**	**0.001834464**	**0.005387**	**0.001834**	**0.005387**

Table 4. Random walk model with drift.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7.728861	6.230897	1.240409	0.216735	-4.58149	20.03922	-4.58149	20.03922
X Variable 1	**-0.00443**	0.006541	**-0.67665**	0.499656	-0.01735	0.008497	-0.01735	0.008497

Table 5. Random walk model without drift: 1st order integration.

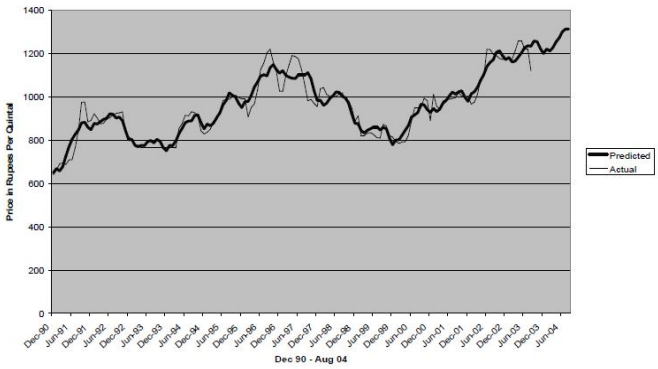
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**0**	**#N/A**	**#N/A**	**#N/A**	**#N/A**	**#N/A**	**#N/A**	**#N/A**
X Variable 1	-0.0476	**0.024222**	-1.96511	**0.051224**	**-0.09545**	**0.000256**	**-0.09545**	**0.000256**

From all the above analysis of the data it can be concluded that the data follows a Autoregression of first order, Integration of first order and no moving average component. Hence, ARIMA (1,1,0) is the best statistical fit for the data.

Step 6: Forecasting deseasonalized data: Forecasting of the deseasonalized data for next 12 data points (one year data) was done on the basis of ARIMA (1,1,0). The Mean square error of the above ARIMA model was 7.88.

Step 7: Adding back seasonal indices: The seasonal index for each month was added back in the forecasted values obtained from the above mentioned process to compute the final forecast. Considering monthly seasonality, the forecasted price is represented by the addition of the seasonality indices with the forecasted deseasonalized data. The graph, comparing the actual and the forecasted values is shown in Fig. 5.

Fig. 5. Forecasted/Predicted compared to actual for 12 more data points.



Step 8: Extrapolation of the model: The model for forecasting was concluded to be ARIMA (1,10) with additive seasonality. The same model was extrapolated to the subsequent eight years of data to compare with the predicted vs. actual and confirm the model.

Step 9: Harmonizing the forecasted data with market reality: The increasing trend in price that was found in the actual data is getting reflected in the forecast also. The seasonal indices came negative for the harvesting months and gradually increase till the lean season persists, which perfectly reflect the market reality. It can be seen that the real data has many sudden fluctuations in price. This may be the result of prevailing political or economic situations like WTO obligations on import or export, Central Govt. election, reduction in import tariff or quota, unforeseen climatic changes, minimum support price (MSP) declared for substitute or competitive oil seeds, change in demand and supply for other substitute oilseeds etc. Hence, it is necessary to keep in mind the ongoing changes in the market conditions in addition to the forecasted prices using this model.

Conclusion

Econometric analysis of the data for the seed prices of soybean helped in understanding the underlying pattern in the data. After analyzing the data of 160 observations, it's apparent that proposed model of ARIMA (1,1,0) with additive seasonality predicts the nature of fluctuation and explains the underlying seasonality. This model can be used by traders, harvesters to minimize the scope for speculation and assume the change in prices of soybean seed for near future. The model can also be used by regulators to predict the future prices and minimize the role of speculators who may otherwise destabilize the market pricing mechanism.

References

1. Gujarati, D.N. and Sangeetha, 2007. Basic Econometrics. Tata McGraw Hill Education Pvt. Ltd. New Delhi, pp.811-854.
2. Soy Protein Council. 1987. Soy Protein Products, Soy Protein Council, Washington, DC.
3. USDA Economic and statistics System. 2000. Soybeans: World supply and demand summary. Retrieved Apr 14 from spectrumcommodities.com/education/commodity/statistics/soybeans.html.