

An Integrated Approach to Forecasting Petrochemical Prices

Authors:

- 1) Dr. Nilotpal Chakravarti, Head, Modeling Lab, DecisionCraft Analytics
- 2) Jai Prakash Mishra, Statistician, DecisionCraft Analytics
- 3) Adityavijay Rathore, Consultant, DecisionCraft Analytics

Contact Information: (n.chakravarti, j.mishra, a.rathore) @decisioncraft.com

Motivation:

The global petrochemical industry trades over \$ 1.9 trillion billion dollars annually. The industry witnesses severe uncertainty in supply and demand, political environment and natural catastrophes that translate into fairly volatile prices. Suppliers, producers and end-users alike, see an increasing share of the uncertainties and risks in business being driven by this volatility of prices.

The industry felt a need for a well-informed view of future petrochemical prices in terms of both value and direction. This would facilitate planning and better hedging against the business risks. It was with a view to cater to the industry's critical need that our client, the industry's leading price reporting and information provider, engaged our services for developing a forecasting product. This forecasting product was conceived with the intention of providing highly accurate price forecasts in the short-term (next 3-4 months) along with a reliable view of medium-term prices (12-month horizon). The forecasting product was to be developed for 5 different product chains across various geographies. This paper will discuss the developed methodology with reference to the Polystyrene chain (Fig 1) for the North American geography comprising of Benzene, Styrene, High Intensity and General Purpose Polystyrene.

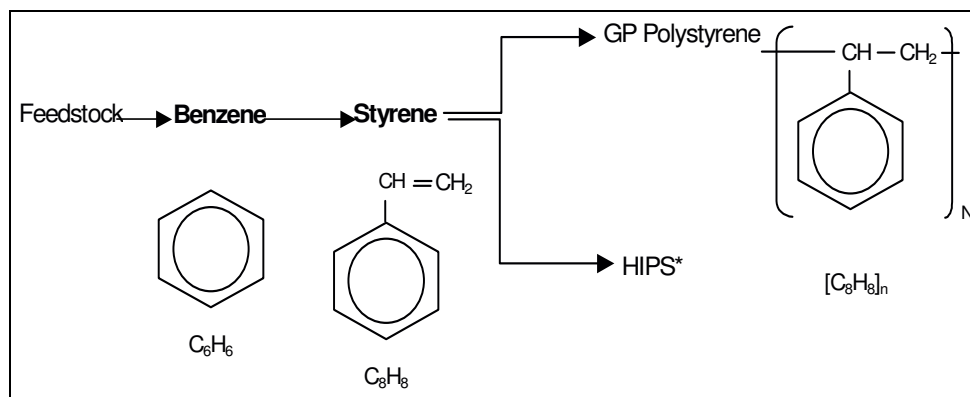


Fig 1: Polystyrene product chain

Approach:

The approach sought to integrate advanced statistical techniques with the domain expertise available with the client in developing the forecasts.

We decided to combine two different modeling approaches:

- 1) Identifying and modeling inherent patterns in the price data and
- 2) Identifying and modeling the relation between important price drivers and prices through a series of expert interviews and statistical correlations using multivariate regression models.

Both approaches will be discussed in detail in the following sub-sections. For clarity, in the discussion below, we focus on a single product, Benzene, which is known for its highly volatile price evolution.

A) Identifying and modeling inherent patterns in the price data

Unadjusted monthly price time series data for the last 12 years was used. Benzene prices have been one of the most volatile with an average 10% change over the previous month and going as high as 35%. The data was tested for 1) trend 2) seasonality and 3) baseline values. The trend element was present in the data but seasonality was absent. There was a shift in baseline values around 2000 and hence, during testing we worked with two different time series: 1) post-1993 and 2) post-2000.

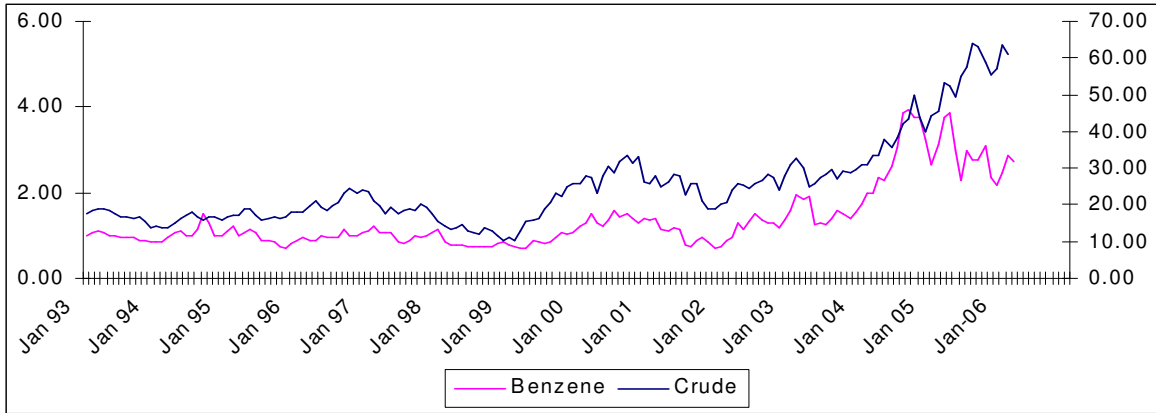


Fig 2: Historical price time series for Benzene & Crude Oil

The project called for a model with long shelf life and a forecasting horizon of 12 months or more. The Auto Regressive Integrated Moving Average (ARIMA) family of models [2] was found to be the most appropriate due to its comprehensiveness. For the ARIMA process, data was tested for stationarity using the Dicky-Fuller test [5][6] at $d = 1$. The ACF and PACF functions for Benzene at $d = 1$ is shown below:

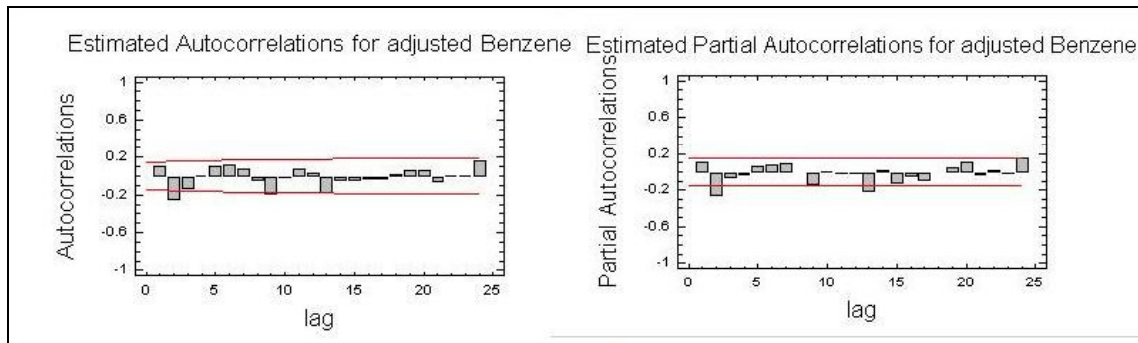


Fig 3: ACF and PACF Functions for Benzene at $d = 1$

ARIMA models with different feasible values of p and q that satisfied the Box-Ljung criteria [7] were further tested according to a rigorous testing schedule over a 12-month validation period. Accuracy for 1-month ahead forecasts was in the range of 86-88% for all the ARIMA models but a potential shortcoming for all the models was their occasional inability to predict directional changes in Benzene prices. Benzene has historically been one of the most volatile price quotes with price swings of over 35% in some months. Interviews with industry experts revealed that the prime reason for the volatility is its proximity to Crude oil in the petrochemical production hierarchy. Experts recommended incorporating Crude oil or Naphtha prices into the model. Hence, ARIMAX (Transfer) [3] models using Crude oil and Naphtha as the exogenous variable were tested. Significant cross-correlation were observed with both Crude oil (0.40) and Naphtha (0.46) at lag 1 i.e. Previous month's Crude prices with current month's Benzene price. Although Naphtha had better cross-correlation with Benzene, Crude oil was chosen as the

final exogenous variable based on lower MAPE (Mean Absolute Percentage Error) and MAE (Mean Absolute Error) over the test period and expert advice. The ARIMAX model with Crude oil resulted in higher accuracy levels along with accurate price direction predictions.

B) Identifying and modeling relation between important price drivers and prices using multivariate regression models

The process of modeling relationships between prices and price drivers was initiated with a series of interviews with industry experts and report editors. These interviews helped us to identify perceived important variables. Sample driver variables considered were classified into 4 main categories: 1) Feedstock prices 2) Industrial indicators 3) Trade indicators and 4) Consumer indicators and included variables such as Consumer Price Index, Consumer Confidence index, Export/Import Price Index, Industrial Production and Producer Price Index.

From a comprehensive list of possibly relevant variables, several were eliminated based on a) Data source reliability and b) Frequency of Data availability (monthly/quarterly/annual). The final list of driver variables was then used to develop multivariate regression models. The variables to be included in the final model were determined on the basis of their statistical significance [1][8] and expert input.

The multivariate regression models were developed mainly with a view to forecasting over 3 to 12 months horizon. While the variable selection process was carried out at lag 0, driver data, especially the macroeconomic indicators, are available only till 2-3 months earlier than the price data. For example: If the petrochemical price data were available till the month of September 2006, driver data would be available till July 2006. Thus, in order to forecast for the period October 2006 – September 2007, regression models can be used in one of the following ways:

- 1) Develop forecasts for driver data till September 2007 and use them as input regression models at lag 0 to generate forecasts.
- 2) Use regression models at lags 3,6,9 and 12 and use them to generate forecasts till July 2007.

Option (2) was chosen, as it would ensure that forecasts were generated only using actual driver variable data. The forecasts were generated using lagged regression as illustrated in the Table 1:

Lag	Regression Data		Forecast Data	
	Price	Driver variable	Forecast Price	Driver variable
6	Upto September-05	Upto March-05	January-06	July-05
9	Upto September-05	Upto December-04	April-06	July-05
12	Upto September-05	Upto September-04	July-06	July-05

Table 1: Illustration of Lagged Regression

The combination of forecasts from lag 3, 6, 9 and 12 regression models was evaluated over a 12-month validation period (Fig 4) resulting in accuracies of 80-85%.

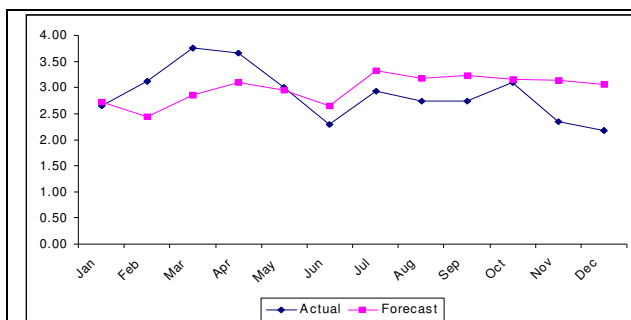


Fig 4: Lagged Regression Forecasts

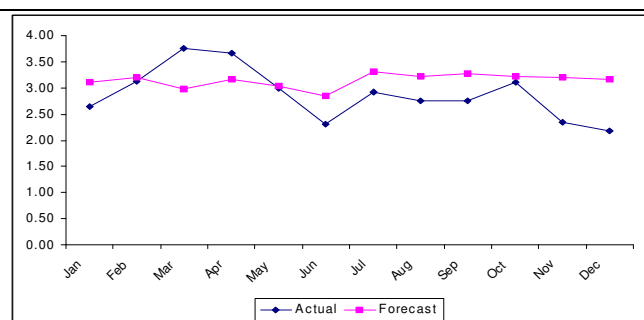


Fig 5: Combined Univariate & Multivariate Forecasts



The final forecasting output was a weighted combination of the forecasts generated by the univariate and lagged regression models (Fig 5). The weights for the univariate and multivariate models were calculated by minimizing the error of the combined forecasts over validation data using Excel Solver. Multivariate models are given a very low weight for short-term forecasts and a higher weight for medium term forecasts. The tools used for time-series modeling were SAS (9.1.3) and Statgraphics.

Conclusion:

The forecasts for Benzene prices generated by the integrated model have been exceptionally accurate over the last 6 months of actual market performance, with a 1-month ahead accuracy of 95%. Models generated for other products have also performed exceptionally with accuracies in the range of 95-99%.

Model development was based on an integrated methodology that incorporates statistical theory, relevant domain expertise, rigorous testing process and implementation practicalities. Forecasts are now being generated on a monthly basis for 5 petrochemical product chains. However this list is expected to grow.

References:

- 1) Bowerman, B., O'Connell, R. and Koehler, A. (2004), *Forecasting, Time Series and Regression*, Duxbury Press, 4th Edition
- 2) Box, G. E. P., and G. M. Jenkins. (1976), *Time Series Analysis: Forecasting and Control*. Rev. ed. Oakland: Holden-Day.
- 3) Box, G.E.P. and Tiao, G.C. (1975), "Intervention Analysis with Applications to Economic and Environmental Problems," *JASA*, 70, 70-79.
- 4) Choi, ByoungSeon (1992), *ARMA Model Identification*, New York: Springer-Verlag, 129-132.
- 5) Dickey, D.A., Hasza, D.P., and Fuller, W.A. (1984), "Testing for Unit Roots in Seasonal Time Series," *Journal of the American Statistical Association*, 79 (386), 355-367.
- 6) Dickey, D.A., and Fuller, W.A. (1979), "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, 74 (366), 427-431.
- 7) Ljung, G.M. and Box, G.E.P. (1978), "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 65, 297-303.
- 8) Neter J., Wasserman W., and Kutner M. H. Kutner (1985). *Applied Linear Statistical Models*. Irwin, Homewood, IL.
- 9) Pankratz, Alan (1983), *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*, New York: John Wiley & Sons, Inc.
- 10) Pankratz, Alan (1991), *Forecasting with Dynamic Regression Models*, New York: John Wiley & Sons, Inc.