

## Data Mining In Oil Price Time Series Analysis

### Daoping Wang

State Key Laboratory of Intelligent Technology and Systems, Tsinghua University,

Beijing 100084, China Tel: 86-10-62333842 Fax: 86-10-62333582

[dpwang@cee.ustb.edu.cn](mailto:dpwang@cee.ustb.edu.cn)

### Litian Cao

Tel: 86-10-62394377 [caolitian04@sina.com.cn](mailto:caolitian04@sina.com.cn)

### Xuedong Gao

Tel: 86-10-62332269 [gaoxuedong@manage.ustb.edu.cn](mailto:gaoxuedong@manage.ustb.edu.cn)

### Tieke Li

Tel: 86-10-62333733 [tieke@public.bta.net.cn](mailto:tieke@public.bta.net.cn)

School of Management, University of Science and Technology,

Beijing 100083, China

### ABSTRACT

*This paper sums up the applications of statistic models such as ARCH-family models, cointegration theory and Granger causality etc in oil price time series analysis and introduces the method of data mining combined with statistic knowledge to analysis oil price time series. In addition, the paper also explains advantages, functions, relevant technologies of this method and its potential applications in hedging the oil shock risk.*

*KEY WORDS: data mining, time series, oil price time series, time-series database*

### INTRODUCTION

In recent years, various kinds of advanced database systems have emerged in response to meet the requirements of new database applications one of which is to handle time-related data (such as historical records or stock exchange data). Therefore, temporal and time-series database systems are developed which are both store time-related data. A temporal database usually stores relational data that include time-related attributes. These attributes may involve several time stamps, each having different semantics. A time series database stores sequences of values that change with time. Data mining techniques may be used to find the characteristics of object evolution or the trend of changes for objects in database. On the other hand, frequent fluctuation of oil price in the international market has aroused our attentions and many models, which are all based on some hypotheses and hence have certain limitations, are built to analyses and forecast the fluctuation. However, for oil price time series, a large amount of disperse and high frequent data information, data mining, combined with these models, may play its technique advantages to effectively organized and manage these data information, mining undiscovered and underlying rules, take out concealed useful information and find out the associate patterns among different time series, which provides credible base for analysis and forecast of oil price risk.

## RELATED STATISTICAL MODELS

Model methods that are often applied to analysis oil price time series include the ARCH model brought forward by Engle in 1982 and thereafter ARCH-family models, Co-integration theory by Engle and Grange in 1987, Granger causality method by Granger in 1969, VAR method by Group of Thirty in 1993 and fractal theory emerging in 1980s.

### *ARCH and ARCH-Family Models*

ARCH effect refers to the clustering phenomena of time series fluctuation. Clustering of a time series indicates the former fluctuation has positive and increasingly reduced effect on that in the future market. Feng C. S., Wu J. C. & Jiang F. (2003) tested the ARCH effect for monthly data of average oil price return ( $R_t = P_t/P_{t-1}$ ) in international market. Feng C. S. (2003) concluded sequence is nonstationary, introduced t-distribution by GARCH and EGARCH models and deduced the parameter estimation of oil price income. Pan H. F. & Zhang J. S. (2005) made analysis on logarithm of domestic oil price return ratio ( $Y_t = 100 \ln(P_t/P_{t-1})$ ) showing time series is stationary and there exists leverage effect of price in oil market.

ARCH model:

Supposed  $\varepsilon_t = \sqrt{h_t} \cdot \nu_t$  is a stochastic progress and  $[\nu_t]$  is independent and synchronous and

$$E(\nu_t) = 0, D(\nu_t) = 1$$

$$h_t = \alpha_0 + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 \quad (1)$$

Where:

$$\alpha_0 > 0, \alpha_i \geq 0, h_t > 0$$

GARCH model :

$$h_t = \alpha_0 + \sum_{j=1}^p \rho_j h_{t-j} + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 \quad (2)$$

Where:

$$\alpha_0 > 0, \rho_i \geq 0, \alpha_j \geq 0, h_t > 0$$

EGARCH model :

$$\ln h_t = \alpha_0 + \sum_{i=1}^{\infty} \pi_i \left\{ \nu_{t-i} - E|\nu_{t-i}| + g^* \nu_{t-i} \right\} \quad (3)$$

### *Co-integration and Causality Analysis*

Co-integration relation indicates the long-term equipoise relation among stable time series that can be expressed by building up co-integration equation while the Granger causality relation can describe the short-term factors having effect on dependent variables of some time series among which there exists co-integration by setting up Error Correction Model (ECM). Yu W. B., Fan Y., Wei Y. M. & Jiao J. (2004) approved the co-integration relation of crude oil prices between in Brent oil futures market and spot market, tested root unit and built up VEC model X to provide new method for hedging and forecasting oil shock. Li X. Y., Wang J. & Gao L. Y. (2005) studied the co-integration relation test of oil price time series in Brent oil and Daqing oil of our country, built up ECM, and presented the causality relation of the two type of crude oil. In addition, Jiao J. L., Fan Y. & Zhang J. T. & Wei Y. M. (2004) also conducted research in the above problems.

Co-integration relation testing:

The general co-integration analysis is orderly involved in the following processes: unit root testing, co-integration testing and Engel-Granger causality analysis.

For two nonstationary sequences,  $x_t$  and  $y_t$ , if they can be transformed stationary sequences after difference, we may determined if there exists co-integration relation between them based on testing if the remnant  $\varepsilon_t$  of the below co-integration regression equation are stationary.

$$x_t = \alpha + \beta y_t + \varepsilon_t \tag{4}$$

Causality relation testing:

If variable x is helpful in improving the precision of forecasting variable y, we will conclude that there exists causality relation between x and y. For the regression models involved in two variables:

$$y_t = \alpha_0 + \sum_{i=1}^m \alpha_i y_{t-ia} + \sum_{i=1}^m \beta_i x_{t-i} + \varepsilon_t \tag{5}$$

$$x_t = \alpha_0 + \sum_{i=1}^m \alpha_i x_{t-ia} + \sum_{i=1}^m \beta_i y_{t-i} + \varepsilon_t \tag{6}$$

We may test the equation:  $\beta_i = 0$ . If we reject the original hypothesis:  $\beta_i = 0$ , then we can conclude that variable x have causality relation with variable y.

*VAR Model*

VAR can be calculated to measure the risk. David Cahedo, Moya & Ismael (2003) applied the historic simulation method of VAR to study on the risk of international oil price.

Basically, there are three types of methods to calculate VAR: non-parameter method (such as historic simulation method), parameter method (such as GARCH method) and extreme value theory.

For a general GDP distribution, the distribution function is:

$$F_u(x) = G_{(\xi, \beta)}(x) = \begin{cases} 1 - (1 + \xi x / \beta)^{-1/\xi} & \xi \neq 0 \\ 1 - \exp(-x / \beta) & \xi = 0 \end{cases} \tag{7}$$

So, the tail estimated value is:

$$\hat{F}(x) = 1 - \frac{N_u}{n} \left(1 + \hat{\xi} \frac{x - u}{\hat{\beta}}\right)^{-1/\hat{\xi}} \quad (8)$$

Based on the above equation, given a confidence  $q$ , VAR can be calculated in reverse:

$$\overline{\text{VAR}}_q = u + \frac{\hat{\xi}}{\hat{\beta}} \left( \left( \frac{N_u}{n} (1 - q) \right)^{-\frac{n}{\hat{\xi}}} - 1 \right) \quad (9)$$

Where:  $N_u/n$  is an experience value by historic simulation method.

In addition, there are other methods in time series analysis such as fractal theory that can discover useful structure underlying complex data system. In recent years, this method is also applied to forecast oil price, for example: Fu L. H. applied it to forecast average yearly price of international crude oil before 2010 in 2004.

Therefore, the present relevant literatures about oil price time series analysis just set up some models which are all built upon some unpractical hypothesis and not provide some underlying relationships and rules of time series data, which results in reducing the accuracy and take on some limitations to a certain extent.

### DATA MINING TECHNOLOGY IN OIL PRICE SERIES ANALYSIS

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid information such as patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data and it also includes analysis and prediction. However, for some historical information involved in millions of data it is quite difficult to analyze. If we apply the data mining tools, combined with statistical methods, we may find out some useful information from enormous data and forecast future data to make correct decisions.

So, combining statistical theory and data mining technology makes oil price analyze more accurate and applicable and provides a new method to hedge oil price fluctuation risk. Moreover, by further study data mining technology in time series, we will also effectively organize and manage abundant, dynamic and various data relating with and affecting oil price fluctuation to analyze part, potential, dynamic and intricate logical relationship, find out influential factors of oil price and make decisions more exactly.

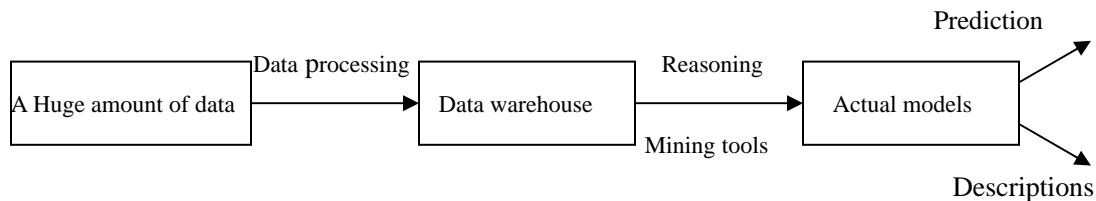
#### *Advantages of Data Mining Combined with Statistical Models*

The above models are based on certain hypotheses. They just reflect the holistic characters of time series, do not exhibit some partial, detailed and hidden characters and thus decrease the accuracy of analysis and forecast. In comparison, we may set up data warehouse of oil price decision analysis based on collecting a huge amount of relevant dynamic data of oil price and relevant information. Then these data and information can be effectively organized and managed to effectively support these data tools of data mining and OLAP. Thus, we can find out long-term and short-term rules influencing oil price underlying time series in data warehouse and carry out dynamic risk decisions according to the useful patterns, relations and information. Data mining involved in many types of data, including nonnumeric value data such as assorted data, and by neural network and decision tree technology can

deal with quite effectively them; the objectives of data mining is a huge amount of data, so data mining have good capacity to express the logic relations among complex variables; data mining technology can continue to dynamically update models after inputting new data according to time and economical environment changes. These characters are all supplements for traditional statistic models.

### *Basic Methods and Relevant Technologies*

Through dividing the sequence, take out characters of each sub-sequence and carry out cluster in these characters and find out several patterns. Then transform pattern to symbol and adopt sequent pattern-generating algorithm to define association rules. By the typical methods we may find out the association relationship of oil prices between in Brent crude oil futures market and other crude oil futures market.



**Figure 1: The General Process of Data Mining**

The data mining technologies that are mainly involved in pattern models and algorithms mining are the difficulty and importance of oil price time series analysis system. By now, some data mining methods have been applied such as: classification knowledge discovery, data congregation, data cluster, association discovery, sequence pattern discovery and trend prediction. At the same time, plenty of data are input into data warehouse and they must be assisted with kinds of analysis tools such as OLAP, statistic and inquiring optimized tools. The capability of these tools has important effects on management decisions. At present, there are intelligent data analysis tools with powerful functions and they not only can collect and deal with data in the process of program running but also be compatible of collecting data of other business system.

In association patterns, Lan Q. J., Ma C. Q., Wu J. H. & Gan G. J. (2004) brought forward a method to discovery changing association patterns in stock market. This method may generate association rules among different time series. The system includes mainly the following parts based on the procedures: sequence division, pattern cluster and association rule mining. And the association rule mining is based on the common effect mechanism idea, that is some common mechanism must exist among time series or there are some causalities between the two series if association pattern frequently is found. This idea may be applied to study price change relation among different oil productions. In addition, Li Bin researched and improved the mining algorithms of time series and made some achievements.

### *The Function of Time Series Mining*

Mining time series includes trend analysis, similarity search, and sequential patterns and periodic patterns. Trend analysis involved in long term or trend movements, cyclic movements, seasonal movements and irregular movements; the similarity search includes a similarity search and subsequence matching; and mining periodicity include mining full periodic patterns, partial periodic patterns and periodic association rules.

## CONCLUSIONS

Data mining technology has some advantages and we may carry out effective forecast and analysis of abundant of oil price time series and relevant data such as production, hedging funds amount and speculative money amount in oil futures and spot markets by data mining combined with time series models to prevent national economic and oil industry from suffering from the oil price shock in market. Moreover, through data mining and model of time series data, we may also analyze the mutual relations among different oil futures and spot markets, find out the relation of price change between crude oil and gasoline and generate the relation among the futures prices and reserves of crude oil and other factors.

## REFERENCES

David Cabedo, Moya & Ismael (2003). Estimating Oil Price 'Value at Risk' Using the Historical Simulation Approach. *Energy Economics*, 2003(25), 239-254.

Feng C. S., Wu J. C. & Jiang F. (2003). ARCH Effect Analysis of International Oil Market. *Journal of the University of Petroleum, China (Edition of Social Sciences)*, 19 (2), 18-20.

Jiao J. L., Fan Y. & Zhang J. T. & Wei Y. M. (2004). The relations between Crude Oil Prices in Domestic Market and in International Market. *Management Review*, 16 (7), 49-55.

Lan Q. J., Ma C. Q., Wu J. H. & Gan G. J. (2004). Data Mining Technology and Its Application and Prosperity in Finance. *Management Review*, 15 (5), 57-62.

Li X. Y., Wang J. & Gao L. Y. (2005). The Relations between Oil Prices in Domestic Market and in International Market. *Statistics and Decision*, 2005 (20), 73-75.

Pan H. F. & Zhang J. S. (2005). Analysis of Domestic Oil price Fluctuation Based on ARCH Models. *Statistical Research*, 2005(4), 16-20.

Yu W. B., Fan Y., Wei Y. M. & Jiao J. (2004). The Cointegration Analysis of Brent Oil Futures Market. *Statistic and Management*, 23(5), 26-32.

(Note: this paper is funded by the open project funds of State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, P R China)