Analyzing Price Movement of Crude Oil through Historical Price Data Distribution by Using Apriori Data Mining Algorithm

Kwan-Hua Sim, Nicholas Ching-Yun Bong, and Kwan-Yong Sim

***Abstract*—Volatile crude oil prices have drawn serious attention lately due to its enormous impact on both economically and politically stability of every oil producing countries in the world. The recent severe plunge of crude oil has immensely tampered the economy of countries that rely excessively on the export of crude oil and natural gas. Highly fluctuated crude oil price has been a major challenge haunting not only businesses, but also governmental agencies in their decision making process concerning risk management and mitigation against possible severe price fluctuation. Therefore it is imperative for exceptional price volatility of crude oil to be studied since conventional financial time series analysis and modeling techniques are inadequate in handling exceptional price volatility. This paper presents a historical crude oil price data distribution analysis by mining the association rules between the characteristic of price distribution and the subsequent maximum price movement. Experiment was conducted on the historical price data of crude oil futures for the period of thirty years to explore the possible association rules by using Apriori data mining algorithm. Evaluation and analysis are performed on the best rules minded to scrutinize each characteristic range of historical price distribution and the maximum future price movement. The outcomes of the experiments reveal a convincing level of association between higher and average downward price movements of crude oil with the historical price distribution that demonstrates positive skewness value. This study institutes a new way of analyzing historical price data to gain information and insight from the distribution of historical data set; the finding stimulates an innovative way on how price data can be interpreted to derive information that is associated to the future price movement of crude oil.**

***Index Terms*—Data mining, financial time series, price distribution, statistical analysis.**

I. Introduction

Financial time series analysis plays a crucial role in the decision making process of individuals and organizations that involve in the transaction of volatile financial instruments such as crude oil. Recent severe price volatility in crude oil has enormously impacted not only companies in oil and gas industry, but also jeopardized the economy of oil producing countries.

The challenge of having an efficient mechanism in handling crude oil price fluctuation has often caused the policy makers get trapped and react passively in realigning policy to ameliorate the sustain volatility of crude oil price. The ripple effect could lead to budget shortfalls, decline in foreign reserve and weaken of currency, which eventually affecting many of the national development programs and public welfare.

The after-shock of recent crude oil price break-down has clearly revealed the inadequacy of convention economic analysis theories in withstanding such high magnitude price volatility, hence noticeable efforts should be devoted to explore a complementary mechanism to aid the trading of financial instruments with high volatility such as light crude oil.

This paper aims to study the future price movement of crude oil by mining the distribution of its historical price data. Apriori data mining algorithm will be employed to mine the rules that are associated to the characteristic of price data distribution and the future price movement. It intends to derive descriptive information from the historical price data to reflex the current state of price, thus infer the possible further price movement of crude oil.

The motivation of conducting experiments and data set used behind this study is further driven by the fact that crude oil futures contract is one of the most traded commodities around the world. It is also the reference point for other commodity indices that are widely traded worldwide, with direct implications on the management of oil contingent claims as well as for risk management activities [1].

Thus, the finding is expected to contribute toward the analysis of historical price data of crude oil by introducing a novel price distribution analytical approach, which intends to reveal the density of transactions that were previously recorded.

This paper begins with Section I as an introduction; Section II concerns the background; Section III elaborates on the employment of Apriori data mining algorithm on crude oil price data; Section IV describes the experiment conducted; at the same time discusses analytical results, and Section V presents the conclusion.

II. Background

Financial time series is an observation of transactions made on a financial instrument chronologically, with numerical and continuous nature, and the value is often considered as a whole instead of individual numerical interval [2]. It is one of the classes of data objects that are

Manuscript received February 23, 2017; revised June 7, 2017. This research is funded by Fundamental Research Grant Scheme (FRGS), Grant No: FRGS/1/2015/ICT04/SWIN/02/1.

Kwan-Hua Sim, Nicholas Ching-Yun Bong, and Kwan-Yong Sim are with Swinburne University of Technology Sarawak Campus, Jalan Simpang Tiga, Kuching 93350 Malaysia (e-mail: khsim@swinburne.edu.my, 4304683@students.swinburne.edu.my, ksim@swinburne.edu.my).

DOI: 10.7763/IJCTE.2017.V9.1162

widely available from financial charting applications.

In pure volatility dependence process, price data of a financial time series is always independent from the past, such process does not make directional forecasting possible since it is closely connected to the size of the price steps in a given time period [3].

However, the feedback system according to the comprehensive Dynamic Financial Market Model is hypnotized to be able to cause change in price distribution of financial time series. This feedback deviates the value of probability of the next price step from random step movements, resulting in non-normal price distribution [3].

Many models have been explored over the years to analysis and model financial time series though financial time series is considered as one of the toughest time series to model [4]. Among the well-known time series stochastic models are Autoregressive Moving Average (ARMA) and Generalized Autoregressive Conditional Heteroskedasticy (GARCH). Though these models possess solid theoretical foundation, their performance in out-sample forecasting on financial time series data remains ambiguous [1].

Notable studies were done by Kang and Yoon, Nomikos and Pouliasis which examined the forecasting of price volatility for daily data of time series front-month energy futures by using a number of stochastic models revealed that these models generally perform well in-sample, but they do not generate any sustainable outstanding performance with out-of-sample data [1].

Nevertheless, the evolution of computational power and big data technology since the last decade has sparked a renewed interest in the area of financial time series alongside with a great deal of research and development attempts in the field of data mining [2].

Various data mining techniques were explored in academic literature; these techniques include pattern discovery and clustering, classification, rule mining and summarization. In this study, focus will be given to rule mining on historical price data of crude oil to discover the association between price data distribution and the future price movement.

Association rule mining is one of the most important techniques in the field of data mining. Association rule mining finds frequent patterns, associations, correlations, or causal structures among set of items or object in a given data set. However, the main fundamental challenge of employing association rule mining in the context of time series data is that it only deals with symbolic and discrete items present in transactions. Hence, time series data in this study will be discretized into segments and converted into symbol before Apriori rule mining algorithm is applied to discover the hidden rules.

The study in this paper is geared toward the effort of exploring the characteristic of historical crude oil price data distribution in relation to the possible future price movement. Price data record and reflect every single transaction done between buyer and seller; hence information revealing the position of participants who were engaged in the market can be mined by scrutinizing the characteristic of historical price distribution collectively [5].

Several noteworthy literatures have been done on skewness and kurtosis measurements of financial time series data. One example of these was done by Kim and White in 2003, they tested the robustness of the skewness and kurtosis measurement on S&P 500 index data [6]. In this study, they concluded that the traditional assumption which stock markets possess negative skewness and severe excess kurtosis might not be always true. In 2005, Bai and Ng tested the skewness, kurtosis and normality for financial time series observations; they revealed that the primary contributor in the normality test is skewness [7].

Another noteworthy contribution by Jondeau and Rockinger focused on the existence and persistence of conditional skewness and kurtosis of various financial series taken at the daily price data. They investigated the existence and persistence of moments in a time-series context. The outcome has drawn the evidence that skewness appears to exist most of the time with strong persistency, and the dynamics of skewness is straight forward to interpret [8].

Although many academic literatures have suggested the existence of skewness in financial time series and the estimation normality, exploration on its impact on future price movements is rather limited. Therefore, this paper will focus on mining the skewness of price distribution from the historical price data of crude oil to discover its association with future price movement.

III. Mining Price Distribution

A. Price Distribution

Price distribution is generally assumed to be random walk with a stochastic diffusion coefficient and a given average for the standard deviation. The most frequent occurrences are at the price where the supply and demand are assumed to be balance. The measures of central tendency and the manner in which prices are scattered around the center point are typically done through descriptive mathematical approach [9].

Apparently, price distribution reflects the dispersion of prices of financial time series data. Information related to the past transactions which have been made, and also where prices were mostly concentrated can be mined from the historical price data distribution.

Fig. 1 illustrates the price distribution of price data over a given interval. The descriptive measurement of standard deviation base on central limit theorem derives crucial information describing the percentage of transactions done in the context of transaction history [9].

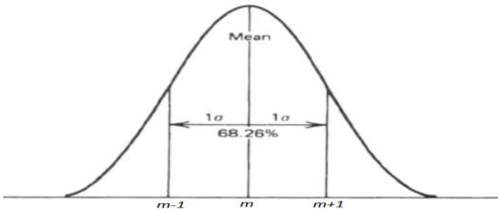


Fig. 1. Price distribution of a given data set.

The predominant measurement in measuring the characteristic of a distribution is referred as the skewness of distribution. The general form of skewness can be stated as:

$$\frac{n\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)(n-2)s^3} \quad (1)$$

where x-bar is mean, s is the standard deviation, and $n > 2$. The relationship of price versus time can be measured as skewness, it denotes the amount of distortion from a symmetric distribution which makes the curve skews to one side and extended on the other.

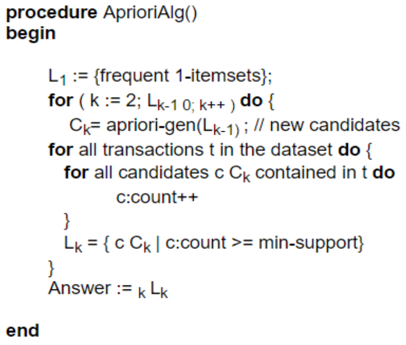
As one of the key properties in measuring normality of data distribution, skewness measures the symmetry of a given data set [10]. A negative value of skewness denotes the distribution of data that leans towards the right. In the context financial time series data, it means the data is skewed towards higher prices. Conversely, positive skewness value is an indication that the price data is leaning towards the lower prices of the price distribution.

According to a classical study done by Sherry in 1992 on stationarity, dependence and randomness of financial time series, few of the important statistical characteristics of financial time series data has been discovered. It has revealed that past prices have an impact on future prices for a given temporal time frame [11]. It proves that financial market, or rather the market participants have memory, though it is not definite, but it is possible to be quantified statistically, and it is definitely worth further investigation.

B. Apriori Algorithm

Apriori algorithm is the fundamental algorithm of association rule mining proposed by R. Agrawal and S. Srikant in 1994 [12]. Apriori algorithm employs an iterative approach known as level wise search through the search space, where k-item sets are used to explore (k+1)-item sets. It begins by finding the set of frequent 1-item sets. The set of that contains one item, which satisfies the support threshold, is denoted by L. In each subsequent iteration, it begins with a seed set of item sets found to be large in the previous iteration. This seed set is then used to generate new potentially large item sets, called candidate item sets, and the actual support for these candidate item sets will be counted. At the end of the iterations, item sets that are actually large or frequent will be determined, and they become the seed for the next iteration. Therefore, L is bused to find L!, the set of frequent 2-itemsets, which is used to find L, and so on, until no more frequent k-item sets are found [13].

Fig. 2 illustrates the process of Apriori algorithm to find all frequent item sets. The process makes the item passes over the database multiple times. The algorithm determines the frequent 1-itemsets by counting the item occurrence in the first pass. The following pass known as k consists of two phases. For the first phase, in the (k-1)th pass, the set of all frequent (k-1)-itemsets, L(k-1) are used to generate the candidate item sets C(k) by apriori-gen() function. The two L(k-1) are joined provided both of the items are the same. The second phase then deletes all item sets that are not in L(k-1) yielding C(k) from the join result. The algorithm then scans the whole database. The hash-tree data structure will increase each time it determines a candidates in C(k) falls into the hash-tree data structure. During the final pass, candidates that are frequent, yielding L(k) are scanned in C(k) [14].



```
procedure AprioriAlg()
begin

        $L_1$ := {frequent 1-itemsets};
        for ( k := 2; $L_{k-1}$ 0; k++ ) do {
           $C_k$= apriori-gen($L_{k-1}$) ; // new candidates
        for all transactions t in the dataset do {
          for all candidates c $C_k$ contained in t do
                   c:count++
          }
          $L_k$ = { c $C_k$ | c:count >= min-support}
        }
        Answer := $_k$ $L_k$

end
```

Fig. 2. Apriori algorithm to find all frequent itemsets.

One of the fundamental challenges of data mining in the context of time series data is the representation and indexation of data [2]. Financial time series is a collection of individual transaction record known as tick data, which is volume intense by nature, hence financial time series inherit the complication of representation and indexing issue too. However, financial industry is quite mature in standardizing the representation and indexation of financial time series data across the industry, all the data are available in commercial charting software with standard intervals across thousands of instruments traded at different exchanges around the globe.

Besides, time series data is also characterized by their numerical and continuous nature, thus a preprocessing step of segmentation is required to discretize the data prior to the data mining process [2]. All the experiments in this study are done by using fixed length sliding window to discretize the historical price data of crude oil. The fixed length is set to be at the interval of 60 time periods to assure the size of sample data points used in plotting the price distribution is statistical sufficient [15].

In order to avoid multiple counting issue attaches to sliding window approach, filtering of attributes is applied to discretize data prior to the implementation of Apriori data mining algorithm [2]. This is to ensure that there is only one single observation corresponds to a change of characteristic in the distribution of historical price data.

IV. Experiments and Evaluation

This paper uses daily price data of NYMEX-CME front-month futures contract, since WTI contract strongly remains the most traded commodity futures throughout the world [1]. All the historical data ware obtained from MetaStock XENITH commodity data, and it covers up to the period of more than thirty years from 30/3/1983 to 3/6/2015 with 8076 data points altogether. The duration of thirty years is expected to encompass all possible market conditions and major economic events.

All price data were discretized into a fixed length with sliding window. Every subsequent window is formed by adding in the next data point, while removing the first data point in the previous window. The maximum price movement of both upward and downward direction in each window of timeframe was recorded as observation. However, filtering is also performed to avoid multiple counting of the same skewness characteristic across multiple sliding windows. Hence, only windows with change of skewness characteristic will be retained. This filtering exercise eventually produced 799 observations.

Apart from price data, discretization on the skewness value is also required. Skewness is discretized into various categories as stated in Table I. As a result, there are altogether seven categories of skewness with three categories for positive skewness and three categories for negative skewness. The range of skewness value is divided equally across all the categories.

The total occurrences of each category after the filtering process to remove all the multiple counting between the sliding windows are listed in the Fig. 3.

Besides, maximum price movements within the observed time widow are discretized into five different ranges of categories shown in Table II. For a maximum price movement that fall into a particular percentage range, the conservative end of the percentage boundary is used to calculate the risk over reward ratio for that observation. This is to increase the reliability of the result and also to off-set all the possible trading costs that might have occurred throughout the process.

TABLE I: SKEWNESS CATEGORY

Category of Skewness	Description (skewness value)
Extremely Positive	>1.25
Very Positive	0.75 – 1.25
Moderately Positive	0.25 – 0.75
Neutral	-0.25 – 0.25
Moderately Negative	-0.25 – -0.75
Very Negative	-0.75 – -1.25
Extremely Negative	<-1.25

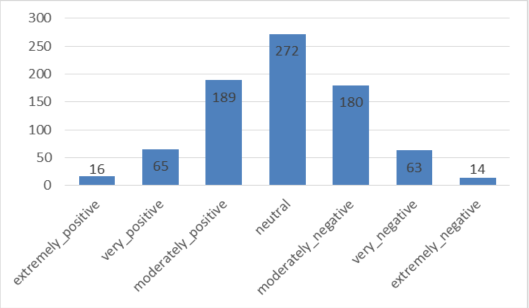


Fig. 3. Occurrences of discretize skewness.

TABLE II: RANGE OF MAXIMUM PRICE MOVEMENT

No.	Highest percentage of price movement
1	0 – <5%
2	5% – <10%
3	10% – <15%
4	15% – <20%
5	>20%

Subsequently, Waikato Environment for Knowledge Analysis (WEKA) software is used to employ Apriori Algorithm in order to mine for possible association rules between the skewness category and the maximum price movement within the observed period.

Experiments were run for both upward and downward movements to emulate the buying and selling activities. Besides, each simulation was tested by adopting stop loss at five percent and ten percent of the entry price to benchmark the risk and reward ratio.

All the best rules with confidence level of above 50% from the experiments with 5% of risk are summarized in Table III and Table IV. All the best rules for both upward and downward price movements have yielded the highest percentage for maximum price movement of 10% to 15%. However, an exceptional maximum downward price movement with more than 20% was discerned for skewness category of moderately positive. Moreover, there were as many as 31 occurrences throughout the experiment that logged an exceptional maximum price movement of more than 20%, given the risk which was controlled at 5%.

TABLE III: UPWARD MOVEMENT WITH 5% RISK

Skewness	Total observa-tion	Highest percentage of price movement	Occurrence	Confidence Level (%)	Risk Reward Ratio
Moderately Positive	65	10% - <15%	38	58	1 : 1.76
Neutral	99	10% - <15%	55	56	1 : 1.86
Moderately Negative	65	10% - <15%	40	62	1: 2.04

TABLE IV: DOWNWARD MOVEMENT WITH 5% RISK

Skewness	Total observa-tion	Highest percentage of price movement	Occurrence	Confidence Level (%)	Risk Reward Ratio
Moderately Positive	58	<20%	31	53	1 : 2.09
Neutral	88	10% - <15%	55	63	1 : 2.11
Moderately Negative	61	10% - <15%	36	59	1 : 1.96

TABLE V: UPWARD MOVEMENT WITH 10% RISK

Skewness	Total observa-tion	Highest percentage of price movement	Occurrence	Confidence Level (%)	Risk Reward Ratio
Moderately Positive	106	5% - >10%	73	69	1 : 1.90
Neutral	157	5% - >10%	109	69	1 : 1.99
Moderately Negative	102	5% - >10%	71	70	1 : 1.85

TABLE VI: DOWNWARD MOVEMENT WITH 10% RISK

Skewness	Total observa-tion	Highest percentage of price movement	Occurrence	Confidence Level (%)	Risk Reward Ratio
Moderately Positive	96	10% - <15%	51	53	1 : 1.99
Neutral	138	10% - <15%	72	52	1 : 2.20
Moderately Negative	96	10% - <15%	50	52	1 : 2.49

As such, distribution of historical data that concentrated slightly to the lower price cluster signifies a possible price downturn which could reach as many as 20% or more.

On the other hand, Table V and Table VI outline the best rules for price movement with confidence level of 50% and above at the risk of 10%. It should be noted that upward price movements produced a lower percentage of maximum price movement across the board with only 10%-15%. Nonetheless, price movements with 10% risk did not yield any higher than average price movement with confidence level of 50% and above that corresponds to the change in characteristic of skewness in historical price data.

It is worth noting that downward price movement with neutral skewness descried a consistently above average risk to reward ratio throughout the whole experiment. Thus, a neutral skewness in the historical prices inclines to the possibility of downward future price movement of crude oil.

However, skewness values that fall outside 0.75 to -0.75 with category other than moderately positive to moderately negative failed to yield any best rules that surpass 50% confidence level.

V. Conclusion

This study has initiated a new epoch of analyzing the distribution of historical price data to mine for the possible associations with the subsequent price movement. It elevates further the analysis and information mining of financial time series data into a new dimension by interpreting the fundamental condition of price concentration area. The outcomes shows some good early signs of positive results, but more in-depth study need to be done on additional attributes in order to further explore for other possible associations that may have existed in the price data. Although experiments were performed in this study to access the relevancy of the proposed approach in analyzing historical price data distribution, there are few key limitations that need to be highlighted. Firstly, only one financial instrument, crude oil futures contract, is used in these experiments. Future work can be done to explore the applications of such analysis technique on other financial instruments, especially those instruments with high degree of liquidity. Secondly, all experiments in this study were done by using fixed sliding window with static time period. Other time periods should be explored in future alongside with a more advanced method in discretizing time series data. Thirdly, this study only focuses on the skewness of the prices distribution and does not include kurtusis in the measurement of price distribution characteristic. The coefficient between skewness and kurtusis in measuring a distribution is a research topic of by itself. Last but not least, future research should also explore the possibility of extending this concept of mining historical price data into a standalone technical indicator, with the expectation to aid the analysis process in financial decision making routine.

References

[1] S. Beniot, "Forecasting the volatility of crude oil futures using intraday data," *Journal of Operational Research, Elsevier,* vol. 235, pp. 643-659, 2014.

[2] T. C. Fu, "A review on time series data mining," *Journal of Engineering Applications of Artificial Intelligence,* vol. 24, pp. 164-181, 2011.

[3] B. Stadnik, "The riddle of volatility clusters," *Business: Theory and Practice,* vol. 15, no. 2, pp. 140-148, 2014.

[4] G. Boetticher, "Teaching financial data mining using stocks and futures contracts," *Journal of Systemic, Cybernetics and Informatics,* vol. 3, no. 3, pp. 26-32, 2006.

[5] B. Stadnik. (June 2011). Dynamic financial market model. *SSRN Electronic Journal.* [Online]. Available: https://www.researchgate.net/publication/228315845_Dynamic_Financial_Market_Model_introduction

[6] T. H. Kim and H. White, "On more robust estimation of skewness and kurtosis," *Finance Research Letter 1,* Elsevier, pp. 56-73, 2004.

[7] J. Bai and S. Ng, "Tests for skewness, kurtosis and normality for time series data," *Journal of Business & Economic Statistics,* vol. 23, no. 1, pp. 49-60, 2005.

[8] E. Jondeau and M. Rockinger, "Conditional volatility, skewness, and kurtosis: Existence, persistence and comovements," *Journal of Economic Dynamics & Control, Elsevier,* vol. 27, pp 1699-1737, 2003.

[9] P. J. Kaufman, *Trading Systems and Methods,* New Jersey: John Wiley & Sons., 2013, pp 15-27.

[10] H.-Y., Kim, "Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis," *Restorative Dentistry & Endodontics,* vol. 3, no. 1, pp. 52-54, 2013.

[11] J. Sweeney, *Campaign Trading, Tactics and Strategies to Exploit the Markets,* New Jersey: John Wiley & sons., 1996, pp. 5-20

[12] S. Singla and A. Malik, "Survey on various improved Apriori Algorithms," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 3, no. 11, pp. 8528-8931, 2014.

[13] C. Kaur, "Association rule mining using Apriori algorithm: A survey," *International Journal of Advanced Research in Computer Engineering & Technology,* vol. 2, no. 6, pp. 2081-2084, 2013.

[14] K. P. Joshi, "Analysis of data mining algorithms," *TechReport, UMBC,* March 1997.

[15] K. H. Sim, K. Y. Sim, I. Goh, and Y. C. Tan, "Forecasting price volatility range of crude palm oil by mining the historical data using hybrid range model," *Intelligent Systems and Applications,* vol. 274, Netherlands, IOS Press, pp. 531-540, 2015.

K. H. Sim was born in Kuching in 1975. He received his BCompSci (Hons) from University Malaysia Sabah in 1999 and MSc. (IT) in 2001 from University Malaysia Sarawak. He is currently a lecturer and program coordinator for bachelor of computer science at the School of Engineering, Computing and Science, Swinburne University of Technology Sarawak, Malaysia. His recent publications include "Forecasting price volatility cluster of commodity futures index by using standard deviation with dynamic data sampling based on significant interval mined from historical data", IEEE Press, pp. 758-763, 2014 and "Analysing Price Movements of Crude Oil Futures by Mining of Dynamic Sample Size through Price Distribution of the Historical Data", International Journal of Modeling and Optimization, pp. 2010-3697, 2015. His has keen research interests in financial time series analysis, data mining and statistical analysis. Mr. Sim is also a member of IEEE.

Nicholas Bong was born in Kuching in 1994. He received his bachelor of computer science from Swinburne University of Technology in 2015. He is currently a master of science candidate at Swinburne University of Technology, Sarawak Campus, Malaysia. His research interest is in time series analysis.

K. Y. Sim was born in Kuching in 1976. He received his BEng (Hons) from the National University of Malaysia in 1999, and masters of computer science from University of Malaya, Malaysia in 2001. He received his doctorate of philosophy from Swinburne University of Technology, Melbourne. He is currently a senior lecturer and the associate head for program development and accreditation at the School of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Malaysia. His recent publications include "Incremental Spectrum Cloning Algorithm for Optimization of Spectrum-Based Fault Localization", Contemporary Engineering Sciences, 7, 1649-1655, 2014 and "Eliminating Human Visual Judgment from Testing of Financial Charting Software", Journal of Software, 9, 298-312, 2014. His research interests include software testing and for embedded system testing. Dr. Sim is a member of IEEE and IEEE Computer Society, a Chartered Engineer (CEng) and member of Institution of Engineering and Technology.

Software Design and Development