# Analysing Price Movements of Crude Oil Futures by Mining of Dynamic Sample Size through Price Distribution of the Historical Data

Kwan-Hua Sim, Isaac Goh, and Kwan-Yong Sim

***Abstract*—Volatile crude oil prices have been drawing a lot of attention lately since it plays a significant role in the world economy. The recent severe price movement has immensely impacted the economy of countries that rely heavily on the production of crude oil and natural gas. While businesses have been struggling in making financial decision to hedge their risk against possible future price fluctuation, governmental bodies and policy makers often caught in the midst of severe volatility. Hence, this paper presents a historical price data distribution analysis by using dynamic data sampling base on the characteristic of the price data distribution. Experiment was conducted on the historical price data of crude oil futures for the period of thirty years. The outcome of the experiment indicates a promising performance demonstrating the relevancy of the proposed approach.**

***Index Terms*—Data mining, data analysis, knowledge discovery, time series analysis, statistical analysis.**

## I. Introduction

Recent high magnitude movements of crude oil price have immensely impacted the economy of Malaysia. As a nation that relies substantially on the revenue from crude oil production, a significant fall in crude oil price has resulted in the loss of essential revenue to the country, which leads to budget shortfalls and the decline in foreign reserve. The ripple effect could also be damaging and long-lasting since it will definitely slow down the development and expansion of oil and gas industry, this may include future oil explorations, investments of oil-related technologies and may potentially lead to other social and political ramifications.

Although conventional fundamental analysis mechanisms offer no solution in dealing with exceptional price movement that occurred recently. Fundamental analysis relies on economic data that are correlated in a highly complex manner, making financial time series one of the hardest time series to model [1].

Nevertheless, many time series models and analysis approaches have been explored over the years to model the movement of the financial time series. Among all these popular techniques, mathematical stochastic modelling from the field of econometric possesses the strongest academically recognition with solid theoretical foundation. Ironically, these mathematical stochastic models have relatively low adoption rate in the industry, and this is reflected by the absence of such models from the major trading software [2].

Fundamentally, all these modeling techniques and analysis approaches are based on a series of historical data with the aim to mine information that is required to forecast the future price movement. In other words, the past price data are analyzed and modeled to capture the pattern of the historical changes that may exist in the prices data. However, as long as historical data is required to be fed into a model for analysis, it inherits a variable concerning the amount of historical data to be used, and the model will inevitably be exposed to statistical inference effect of data snooping [3].

Although most stochastic models are capable of adapting to the time evolution of price fluctuation, the statistical nature of the mechanism does not imply any causation in the context of price movement. The models are therefore constrained by the sample of data used during the modeling, and the challenge in selecting sample size has undoubtedly tampered its practicality in the real-world application.

Thus, this study aims to propose a price distribution approach with dynamic data sampling of the historical price data, and subsequently forecast the potential price movement. The distribution of price data is quantified by measuring the skewness and kurtosis values of the distribution, and the attributes and characteristics of the price distribution will be evaluated in correspond to the forecasted future price movement. The performance of the proposed approach will be benchmarked against static data sampling.

This paper begins with Section I as an introduction; Section II concerns the background; Section III elaborates on the proposed price distribution approach; Section IV describes the experiment conducted; at the same time discusses analytical results, and Section V presents the conclusion.

## II. Background

One of the key motivations behind this study is that it analyse the price movement on the most traded commodity in the world, and the results may accord the management of oil contingent claims, as well as for risk management and portfolio management activities [4]. Since competitive forecasts of future price movement have direct applications on risk management activities, risk mitigation instruments such as futures and options, hinge on specific properties of price distribution to determine their coast and price.

Though price in financial time series is a value reflecting a particular financial asset, it also encompasses and factors in the consensus of future expectation inferenced by market

Manuscript received September 29, 2015; revised November 30, 2015.

The authors are with Swinburne University of Technology Sarawak, Jalan Simpang Tiga, 93350 Kuching, Sarawak, Malaysia (e-mail: khsim@swinburne.edu.my).

DOI: 10.7763/IJMO.2015.V5.494

participants, and the observed return distribution is the aggregate result of the actions of all the market participants. Thus, the deviation of expectation from the value of the asset at current state leads to the fluctuation of financial price series. Unfortunately, it is nearly impossible to validate the pricing of an asset due to manifold influences that determine the price value [5].

In econometric models, the popular ARIMA (autoregressive integrated moving average) and ARCH (AutoRegressive Conditional Heteroskedasticity) with their numerous extensions were developed to forecast price volatility, and they are extensively studied in academic research. Certain extensions of these models have demonstrated the capability to reproduce many of its observed characteristics such as long memory and leverage effects. Optimal values for the model parameters are normally derived by fitting the observed distribution of prices [6].

In fact, the modeling of the crude oil futures returns volatility has been examined in a number of well known contributions, and most of those models are built on ARIMA and GARCH specification. Sadorsky (2006) provides a through empirical analysis of the modeling and forecasting of crude oil futures, he gathers evidence of good performances of a number of GARCH-type models that are able to out-perform random walk model. Other recent noteworthy contribution were done by Engle [7] in 1995 on ARCH specification, and Witzany [8] in 2013 on price volatility dependence with Markov Chain Monte Carlo sampling algorithm. Although these models embrace statistical compositions of price data distribution, scrutinization on the causation that drives the price movement is still lacking. Another empirical study by Nomios and Poulias [9] outlined that regime shits are present in price data and dominate GARCH effects. Generally all these econometric models seem to proclaim good performance for in-sample data, but they are generally suffered from poor out-of-sample performance [4].

Moreover, most of the financial time series forecasting models to date lean to be oriented towards a single point of time into the future, and this has also tampered significantly the practicality of such models. As such, models with the capability of comprehending many points in time into the future or even a continuous development of prices should be explored [5]. Essentially, it is an ancient financial mathematically theory that a continuous model for development of prices is required within the whole period of observation.

Generally, a Gaussian distribution covers the whole period of observation since prices would depend on the stochastic behavior of a large number of market participants, and it will eventually produce a Gaussian distribution [6]. In particular, price fluctuation is normally measured by a dispersion variable such as variance or standard deviation in the presence of non-normal price distribution [10].

A classical study by Sherry (1992) on stationarity, dependence and randomness of financial time series has addressed few of the important statistical characteristics of financial time series data. The author concluded that for given a certain temporal time frame, past prices have an impact on future prices [11]. It implies 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.

## III. Dynamic Sampling through Price Distribution

Price distribution is generally assumed to be random walk with a stochastic diffusion coefficient and a given average for the standard deviation. Without losing the generality of probability distribution function, it can be assumed in the form of (1).

$$f_w(x) = \frac{g(x/w)}{wN_0} \tag{1}$$

With $N = \int_{-\infty}^{+\infty} g(z)dz$ and $w$, a positive quantity, being the scale of the distribution [6].

Theoretically, the most frequent occurrences are at the price where the supply and demand balances. The measures of central tendency and the manner in which prices are scattered around the center point are done by three typical descriptive measurements known as dispersion, skewness and kurtosis. Standard deviation is the best estimate of distribution by measuring the degree of data dispersion from the mean. It is a form of measuring average deviation from the mean, which uses the root mean square in (2).

$$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}} \tag{2}$$

where the differences between individual prices and the mean are squared to emphasize the significance of extreme values; and the total is scaled back with square root function. The value of one standard deviation represents a clustering of around 68% of the sample data that form the distribution, and about 95% of data fall within two standard deviation away from the mean value [12].

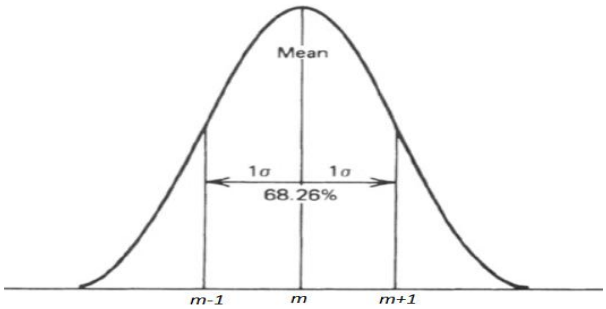



Fig. 1. Price distribution of a given data set.

Fig. 1 illustrates the price distribution of all the price data from a given sample price data over a certain interval. The descriptive measurement of standard deviation base on centre limit theorem derives crucial information in the context of price transaction history. That is, approximately 68% of the transactions were previously traded between price m-1 and price m+1. Hence, an immediate price movement above m+1 indicates that more than 80% of the previous transactions are

below the current price level of m+1, and via versa for immediate price movement below m-1. A significant majority of more than 80% of market participants are in agreement, and that will create a potentially high impact future price movement collectively. This causation effect opens up the possibility of subsequent high probability price movement.

Alike most of the other time series modelling techniques, price distribution analysis still relies on historical price data as the input; hence it is bounded by the variable concerning the number of price data points to be used as sample in plotting the distribution. Therefore, this paper proposes a dynamic data sampling by examining the characteristic of the price distribution from the historical price data. The sample size that will be adopted for price distribution analysis is based on the criteria of the historical price data distribution. In data distribution theory, a close to Gaussian distribution will have a better statistical representation than those distributions with leptokurtic and platykurtic distributions.

The predominant measurements that are widely used to measure the shape of a distribution are known as skewness and kurtosis. The general form of skewness can be stated as:

$$\frac{n\sum_{i=1}^{n}(X_i-\bar{X})^3}{(n-1)(n-2)s^3} \tag{3}$$

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

$$\frac{n(n+1)\sum_{i=1}^{n}(X_i-\bar{X})^4}{(n-1)(n-2)(n-3)s^4}-\frac{3(n-1)^2}{(n-2)(n-3)} \tag{4}$$

where *x*-bar is mean, *s* is the standard deviation, and $n > 3$. Kurtosis describes the peakedness or flatness of a distribution, it is a good assessment on frequency of price movement over certain price levels, and the frequency distribution is accumulated dynamically [12]. Notably, platykurtic distribution is believed to be a common phenomenon financial time series distribution. It is a property of probability distributions that commonly known as fat-tailed distribution; it exhibits an extreme scale at both ends, and demonstrates power of law decomposition.

Indeed, most of the standard deviation based technical indicators in the field of technical analysis simply assume a normal or symmetric distribution from a given fix length of sample size [12]. This will inevitably invite the risk of data snooping since a mechanism of sample size selection is literally absent, it reiterates the gap between pure statistically computation and causation effect in the context of price movement.

Importantly, a frequency distribution with no assumption on the distribution shape, but records the amount of time prices transacted at a specific price level will reveal the density of transactions made previously at a particular price level. The selection of sample size that is determined by the characteristic of the price distribution will be examined in this study.

## IV. Experiments and Evaluation

All the experiments in this study were conducted by using the historical data of crude oil futures. The historical data ware obtained from MetaStock XENITH commodity data, and it covers up to the period of more than thirty years from 30/3/1984 to 31/3/2015 with 8031 data points altogether. The duration of thirty years is expected to encompass all possible market conditions and major economic events.

The experiments were run from the beginning of the experiment period, the value of skewness and kurtosis will be calculated as the subsequent data set is added in. Experiments were conducted for various thresholds of skewness and kurtosis values, observations were recorded whenever a new price data added has exceeded one standard deviation away from the mean, denoting a potential high impact price movement collectively. Forthwith, the skewness and kurtosis values were logged before looping through the next price data point, the magnitude of price movement was also traced throughout the observation period. The observation will reach cessation once the price movement retraced back to the average price level since the beginning of the observation period.

An observation is classified as a valid price movement if there is a minimum price movement of at least 2% during the observation period, or else it will be categorized as an invalid observation. The buffer of 2% is selected to cater the possible transaction costs in real life application, though it generally costs less than 1% in real-world scenario.

A similar experiment was then carried out with fix amount of sample size throughout the experiment period, without any consideration on skewness and kurtosis of the price distribution. The static sample size was selected from 20 up to 120 data period since those are the common data periods used in financial market to reflect average price for both short term and long term prospect by professional financial analysts [12]. Next, the results acquired are dissected and analyzed to scrutinize the characteristics of price distribution in producing valid observations.

As illustrated in Fig. 2 and Fig. 3, price distributions with kurtosis value of 0.6 to 1.0 documented the accuracy of above 80% in forecasting subsequent price movement once the price moves beyond one standard deviation from the mean of the distribution. In fact, Fig. 2 shows that the performance is consistently improved as the value of kurtosis increases, indicating a more obvious characteristic of leptokurtic distribution. This insinuates that leptokurtic distribution with thin-tailed distribution has more significant statistical representation than playtikurtic or fat-tailed distribution, and this is also inlined with generally statistical postulation.

Adversely, normal distributions are generally expected to produce better performance. But, result in Fig. 4 divulges a contrary phenomenon with steady trend of improvement in term of accuracy as the distributions skew positively. On the other hand, Fig. 5 reveal that negative skewness do not demonstrate any obvious impact on the accuracy in forecasting price movement. Nevertheless, the generally accuracy of at least 70% has been obtained by the proposed

dynamic data sampling approach base on skweness thresholds.

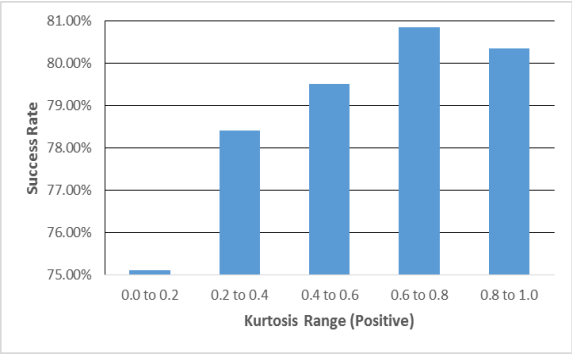



Fig. 2. Price distribution with positive kurtosis value.

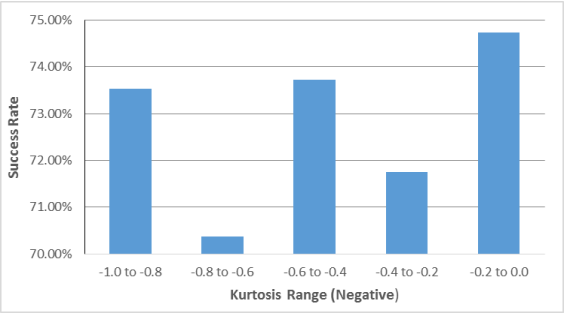



Fig. 3. Price distribution with negative kurtosis value.

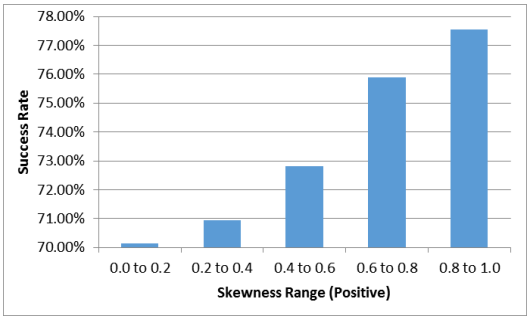



Fig. 4. Price distribution with positive skewness value.

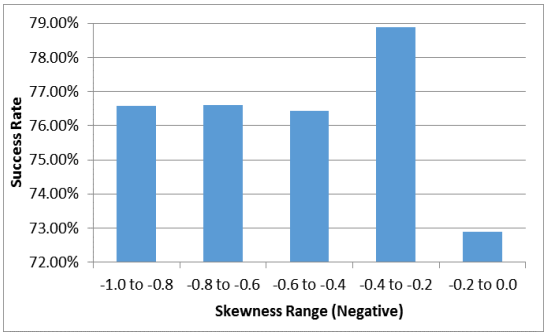



Fig. 5. Price distribution with negative skewness value.

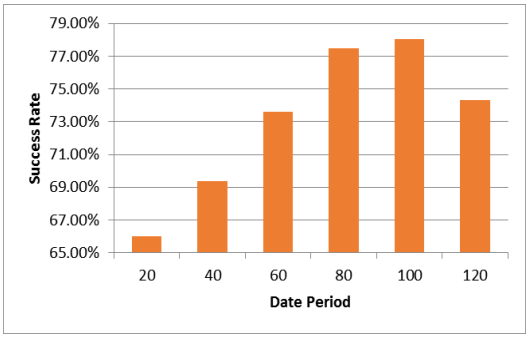



Fig. 6. Price distribution with static sample size.

Meanwhile, Fig. 6 outlines the performance of price distribution with static data sample size. The performance of price distribution with sample data size of above 60 data periods accomplishes a performance that is on par with the best performance registered by the proposed dynamic sampling approach. This has further conformed the common statistical principle that larger amount of data will lead to a more significant result statistically. It indicates that more price data points will accumulate larger collective impact to fuel a subsequent price movement once the price moves beyond 80% of the previously transacted price levels.

TABLE I: SUMMARY OF KURTOSIS THRESHOLDS

| Kurtosis | -1.0 to 0.8 | -0.8 to -0.6 | -0.6 to -0.4 | -0.4 to -0.2 | -0.2 to 0.0 | 0.0 to 0.2 | 0.2 to 0.4 | 0.4 to 0.6 | 0.6 to 0.8 | 0.8 to 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Occurrence** | 306 | 351 | 331 | 308 | 277 | 229 | 290 | 166 | 141 | 112 |
| **Average Maximum Movement (%)** | 8.15 | 7.56 | 8.07 | 8.23 | 8.54 | 9.25 | 10.21 | 11.07 | 10.79 | 11.49 |

TABLE II. SUMMARY OF SKEWNESS THRESHOLDS

| Skewness | -1.0 to -0.8 | -0.8 to -0.6 | -0.6 to -0.4 | -0.4 to -0.2 | -0.2 to 0.0 | 0.0 to 0.2 | 0.2 to 0.4 | 0.4 to 0.6 | 0.6 to 0.8 | 0.8 to 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Occurrence** | 111 | 171 | 229 | 270 | 306 | 308 | 289 | 206 | 141 | 98 |
| **Average Maximum Movement (%)** | 10.67 | 8.98 | 9.07 | 8.72 | 7.97 | 8.18 | 8.35 | 9.29 | 11.78 | 13.62 |

TABLE III: SUMMARY OF STATIC DATA PERIOD

| Data Period | 20 | 40 | 60 | 80 | 100 | 120 |
|---|---|---|---|---|---|---|
| **Occurrence** | 509 | 297 | 193 | 151 | 132 | 113 |
| **Average Maximum Movement (%)** | 6.52 | 8.64 | 11.09 | 12.74 | 13.29 | 14.41 |

Table I to Table III summarize the occurrence of observations throughout the experiment period, and also the average maximum movement among the valid observations. It is interesting to discern in Table I that occurrence reduces

consistently toward leptokurtic distributions. This has conformed to the general belief that financial time series tend to exhibit fat-tail behaviour. Similarly, outcome stated in Table II has also demonstrated that probability distribution saturates near mean value of a given sample data. Interestingly, the highest average maximum price movement is recorded at 120 data periods in static data sampling.

## V. Conclusion and Future Works

This study has initiated a new epoch of how the size of sample data can be mined from the distribution of historical price data in financial time series modelling and analysis. Future works will explore the possibility of extending this approach other financial instruments. Since both skewness and kurtosis are examined and evaluated separately in this study, effort should also be attempted in future to explore away to combine both values together in evaluating a price distribution.

## Acknowledgment

This research is funded by Fundamental Research Grant Scheme (FRGS/2/2013/ICT01/SWIN/03/1)

## References

[1] 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.

[2] K. H. Sim, I. Goh, K. Y. Sim, 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*, pp. 531-540, 2014. IOS Press.

[3] G. Norden, *Technical Analysis and the Active Trader*, New York: McGraw-Hill, 2006, pp. 19-39

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

*[5]* E. Korn and R. Korn, "Modelling stock prices," O*ption Pricing and Portfolio Optimization- Modern Methods Of Financial Mathematics*, Kaiserslautern, Germany: University of Kaiserslautern, Department of Mathematics. pp 152-188, 2011.

[6] R. Bartiromo, *Maximum Entropy Distribution of Stock Price Fluctuations*, Rome: Cornell University Library, 2013.

[7] R. F. Engle and K. F. Kroner, "Multivariate simultaneous generalized ARCH," *Econometric Theory*, 11, pp. 122-150, 1995.

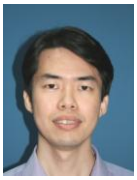
[8] J. Witzany, "Estimating correlated jumps and stochastic volatilities," *Prague Economic Papers 2*, pp. 251–283, 2013.

[9] N. Nomikos and P. Pouliasis, "Forecasting petroleum futures markets volatility: The role of regimes and market conditions," *Economic*, vol. 33, pp. 321-337, 2011.

[10] T. Voituriez, "What explains price volatility changes in commodity markets? Answers from the world palm-oil market," *Agricultural Economics*, vol. 25, pp. 295-301, 2001.

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

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



**K. H. Sim** was born in Kuching, 1975. He received his BCompSci (Hons) from University Malaysia Sabah in 1999 and MSc. (IT) in 2001. 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 research interests include financial time series analysis, data mining and statistical analysis. Mr Sim is also a member of IEEE.

**Isaac Goh** was born in Bintulu, in 1992. He received his bachelor of ICT from Swinburne University of Technology in 2013. 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 B.Eng (Hons) from the National University of Malaysia in 1999 and masters of computer science from University of Malaya, Malaysia in 2001. 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 research interests include software testing and for embedded system testing. Mr. Sim is a member of IEEE and IEEE Computer Society.