Crude Oil Forecasting With An Improved Model Based on Wavelet Transform and Linear Regression Model

Ani Shabri

Abstract—This paper presents a hybrid wavelet linear regression (WLR) model that combines both wavelet technique and the linear regression (LR) model for crude oil forecasting. Based on the purpose, the main time series was decomposed to some multifrequency time series by wavelet theory and these time series were imposed as input data to the LR for forecasting of crude oil series. To assess the effectiveness of this model, daily Brent crude oil prices has been used as the case study. Time series prediction capability performance of the WLR model is compared with the single LR, ARIMA and GARCH models using various statistics measures. As seen in comparison, WLR yielded more accurate than of any individual model and offered a practical solution to the problem in crude oil forecasting.

Keywords—Wavelet, linear regression, GARCH, ARIMA

I. INTRODUCTION

MODELLING and forecasting crude oil price is among the most important issue in the world economy because accurately measuring of crude oil future prices is an important component of the price linkage between spot and future markets. Thus, a better understanding of the dynamic of crude oil futures prices can provide useful to energy researchers, market participants and policymakers. However, accurate time series forecasting is one of the greatest challenges in real-world crude oil prices series because there is a great deal of nonlinearity and irregularity.

In general, the stochastic models such autoregressive moving average (ARIMA) and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) types modeled are widely used for crude oil forecasting ([1],[2],[3],[4],[5]). Usually, the above models can provide good prediction results when the prices series under study is linear or near linear. Numerous experiments have demonstrated that the prediction performance might be very poor if one continued using these traditional statistical and econometric models. The main reason leading to this phenomenon is that the ARIMA or GARCH models are basically linear models assuming data are stationary, and have a limited ability to capture non-stationary and non-linear in the crude oil prices series data.

To remedy the above shortcomings, recently, new hybrid models on wavelet transform process have been improved for forecasting crude oil and obtain the best performances ([6], [7],[8],[9],[10]). They observed the use of wavelet techniques to pre-process time series data into decomposed wavelet coefficients of different decompositions produced significant results than the original time series when used as input. Such hybrid models show significant advantages over the traditional AI models such as neural network, least square support vector machines (LSVM) and fuzzy neural network. This indicates that the wavelet can be a promising tool in the decomposition of the time series.

In this paper, a hybridization of wavelet and linear regression model (WLR) has been proposed to forecast crude oil prices. To prove the application of this model, the daily Brent crude oil is chosen as the case study. Finally, this study is to evaluate the performance of the new model ability, this model was compared with the performance of single LR, ARIMA and GARCH models.

II. THE ARIMA MODEL

Box–Jenkins models are the most common and best-known statistical technique that is used for forecasting time series [11]. Box-Jenkins model are also known as ARIMA ($p,d,q$) models and expressed as:

$$x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

(1)

Where $p$ is the order of the non-seasonal autoregressive, $q$ is the order of the non-seasonal moving average, and $\varepsilon_t$ is the random error. First or second order of differencing is used if the original data is non-stationary.

III. THE GARCH MODEL

The most popular volatility model is standard GARCH (1,1) model proposed by Bollerslev [12] for conditional volatility. The standard GARCH (1,1) can be described as follows

$$r_t = \mu_t + \varepsilon_t = \mu_t + h_t^{1/2} \eta_t \quad \eta_t \sim N(0,1)$$

Ani Shabri is with the Science Mathematic Department, Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor, Malaysia (phone:607-5534274; fax: 607-5566162; e-mail: ani@utm.my ).
\[ h_t = \omega + \alpha e_t^2 + \beta h_{t-1} \]  

(2)

where \( r_t = \ln(x_t / x_{t-1}) \) is log return, \( \mu_t \) is the conditional mean, \( h_t \) is the conditional variances and \( \eta_t \) is a standardized error follows a normal distribution with a mean of zero and variance of one.

IV. AN APPLICATION

The daily Brent crude oil spot price series, which is treated as the benchmark crude oil for international oil markets are used in our analysis. The daily data from May 20, 1987 to September 30, 2006 with a total of 4933 observations. For convenience of WLR modeling, the data from May 20, 1987 to December 31, 2002 is used for the training set (3965 observations), and the remainder is used as the testing set (968 observations). In practice, short-term forecasting results are more useful as they provide timely information for the correction of forecasting value. Fig. 1 shows the daily crude oil prices from January 1, 1986 to September 30, 2006.

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2} \quad \text{and} \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{x}_i| \]

Where \( x_i \) is the actual and \( \hat{x}_i \) is the forecasted value of period \( t \), and \( n \) is the number of total observation.

V. WAVELET TRANSFORM

Wavelet transformations provide useful decomposition of original time series by capturing useful information on various decomposition levels. According to the Mallat’s theory, the original discrete time series \( x(t) \) can be decomposed into a series of linearity independent approximation and detail signals by using the inverse discrete transform is given by [13]

\[ x(t) = T + \sum_{m=1}^{M} \sum_{n=0}^{2^m-1} W_{m,n} 2^{-m/2} \psi(2^{-m} t - n) \]  

(3)

Where \( W_{m,n} = 2^{-m/2} \sum_{j=0}^{N-1} \psi(2^{-m} t - n) x(t) \) is discrete wavelet transformation (DWT), \( \psi(t) \) called the mother wavelet and \( N \) is size sample, \( m \) and \( n \) are integers that control the scale and \( t \). The Eq. (3) can be simplify as

\[ x(t) = A_m(t) + \sum_{m=1}^{M} D_m(t) \]

\( A_m(t) \) is called approximation sub-series or residual term at levels \( M \) and \( D_m(t) \ (m = 1, 2, ..., M) \) are detail sub-series which can capture small features of interpretational value in the data.

The wavelet coefficients and scale coefficients of the daily Brent crude oil level time series derive a Mallat decomposition algorithm using Daubechies wavelet are shown in Fig. 2.

VI. HYBRID WAVELET-REGRESSION MODEL

The coupled wavelet and linear regression (LR) models are LR models which use sub-time series components obtained using DWT on original data as input. Each sub-series component plays a different role in the original time series. However, there is no existing theory to tell how many decomposition levels are needed for any time series. In this study, three wavelet decomposition levels (2-4-8) were employed, as well as in similar studies by Nourani et al. [14].

The coefficient of determination \( (R^2) \) between each D sub-time series and original data were used to determine the effectiveness of wavelet components. The quantity, \( R^2 \) is a measure of the proportion of variability explained by the fitted model. The \( R^2 \) values of D sub-time series and original data are given in Table I. It can be seen that the D1 component \( (R^2 = 0\%) \) shows that there is no variability in data is contributed by the D1. However, the wavelet component D2 and D3 show significantly higher \( R^2 \) compared to the D1. According to the \( R^2 \) analyses, the effective components D2 and D3 were selected as the dominant wavelet components. Afterward, the significant wavelet components D2, D3 and approximation (A3) component were added to each other to constitute the
new series to the LR model. Fig. 2 shows the structure of the WLR model. Fig. 3 shows the original crude oil prices and their Ds, that is the time series of 2-month mode (D1), 4-month mode (D2), 8-month mode (D3), approximate mode (A3), and the combinations of effective details and approximation components mode (A2 + D2 + D3). Five different combinations of the new series input data is considered in this study.

A program code including wavelet toolbox was written in MATLAB language for the development of LR and WLR model. The forecasting performances of the LR and WLR models in terms of the RMSE and MAE at testing phase are compared and shown in Table II. Table II shows the M1 with 1 lags obtained the best RMSE and MAE statistics of 1.0039 and 0.7558, respectively for LR model. However for WLR model, the M5 with lags 5 obtained the best RMSE and MAE statistics of 0.5620 and 0.3810, respectively

For further analysis, the best performance of the LR, WLR, ARIMA and GARCH models were compared with the best results of ARIMA and forward neural network (FNN) studied by Yu et al. [15]. In Table III, it shows that WLR has good performance they outperform LR, ARIMA, GARCH, Yu' ARIMA and Yu' FNN models in terms of all the standard statistical measures. This results show that the new series (DWT) have significant extremely positive effect on regression model results.

Fig. 4 shows the Box-plot for the ARIMA, GARCH, LR and WLR models for testing period. It can be seen that the errors of WLR model quite close to the zero. Overall, it can be concluded the WLR model provided more accurate forecasting results than the other models for crude oil forecasting.

## Table I

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1.0039</td>
<td>0.7558</td>
</tr>
<tr>
<td>M2</td>
<td>1.0078</td>
<td>0.7590</td>
</tr>
<tr>
<td>M3</td>
<td>1.0078</td>
<td>0.7590</td>
</tr>
<tr>
<td>M4</td>
<td>1.0089</td>
<td>0.7604</td>
</tr>
<tr>
<td>M5</td>
<td>1.0089</td>
<td>0.7604</td>
</tr>
</tbody>
</table>

## Table II

<table>
<thead>
<tr>
<th>Model</th>
<th>Lag</th>
<th>R^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>1</td>
<td>0.9918</td>
</tr>
<tr>
<td>M2</td>
<td>2</td>
<td>0.9901</td>
</tr>
<tr>
<td>M3</td>
<td>3</td>
<td>0.9895</td>
</tr>
<tr>
<td>M4</td>
<td>4</td>
<td>0.9895</td>
</tr>
<tr>
<td>M5</td>
<td>5</td>
<td>0.9967</td>
</tr>
</tbody>
</table>

## Table III

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARIMA(7,1,6)</td>
<td>0.9997</td>
<td>0.7479</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>0.9918</td>
<td>0.7425</td>
</tr>
<tr>
<td>LR</td>
<td>1.0039</td>
<td>0.7558</td>
</tr>
<tr>
<td>WLR</td>
<td>0.5620</td>
<td>0.3810</td>
</tr>
<tr>
<td>Yu' ARIMA</td>
<td>1.768</td>
<td>-</td>
</tr>
<tr>
<td>Yu' FNN</td>
<td>0.743</td>
<td>-</td>
</tr>
</tbody>
</table>

## VII. VI Conclusion

This paper proposes the application of the WLR technique for the modelling of crude oil series. The WLR models are obtained by combining two methods, an LR model and discrete wavelet transforms. The series was decomposed at 3 decomposition levels (2–4–8 months). The sum of the effective details and the approximation component were used as inputs to the LR model. The WLR models were trained and tested by applying different input combinations for different crude oil series data. The performance of the proposed WLR model was compared to forecasting using regular LR, ARIMA and GARCH models. Comparison of the results indicated that the WLR model was substantially more accurate than the other models. The study concludes that the forecasting abilities of the LR model in short-term daily crude oil time series are found to be improved when the wavelet transformation technique is adopted for data pre-processing. The decomposed periodic components obtained from the DWT technique are found to be the most effective in yielding an accurate forecast when used as inputs for LR models.

## Acknowledgment

The authors thankfully acknowledged the financial support that afforded by Universiti Teknologi Malaysia under FRGS Grant (VOT 64F088).

## References


Ani Shabri is a senior lecturer at Department of Science Mathematic, Faculty of Science, Universiti Teknologi Malaysia, Skudai, Malaysia. He obtained his Bachelor and Master degrees in Industrial Science (Mathematics) from Universiti Teknologi Malaysia in 1992 and 1995 respectively. Then he received his PhD in Mathematics from Universiti Kebangsaan Malaysia in 2007. He has published several papers on flood frequency analysis, soft computing and time series. His research interests include the development and application of forecasting methods based on neural networks, wavelet, fuzzy time series, support vector machine and other computational intelligence techniques, and also on analysis of flood frequency.