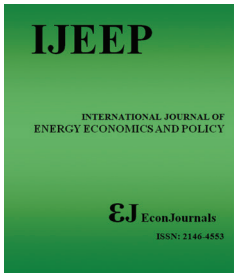


International Journal of Energy Economics and Policy

ISSN: 2146-4553

available at http: www.econjournals.com

International Journal of Energy Economics and Policy, 2019, 9(5), 135-141.

A Hybrid Autoregressive Integrated Moving Average-phGMDH Model to Forecast Crude Oil Price

Richard Manu Nana Yaw Sarpong-Streetor*, Rajalingam A/L Sokkalingam, Mahmoud bin Othman, Dennis Ling Chuan Ching, Hamza bin Sakidin

Fundamental and Applied Science Department, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak Darul Ridzuan, Malaysia. *Email: rsarpongstreetor@gmail.com

Received: 02 April 2019 **Accepted:** 05 July 2019 **DOI:** https://doi.org/10.32479/ijeep.7987

ABSTRACT

Crude oil price fluctuations affect almost every individual and activity on the planet. Forecasting the crude oil price is therefore an important concern especially in economic policy and financial circles as it enables stakeholders estimate crude oil price at a point in time. Autoregressive integrated moving average (ARIMA) has been an effective tool that has been used widely to model time series. Its limitation is the fact that it cannot model nonlinear systems sufficiently. This paper assesses the ability to build a robust forecasting model for the world crude oil price, Brent on the international market using a hybrid of two methods ARIMA and polynomial harmonic group method of data handling. ARIMA methodology is used to model the time series component with constant variance whilst the polynomial harmonic group method of data handling is used to model the harmonic ARIMA model residuals.

Keywords: Autocorrelation, Harmonics, Residuals
JEL Classifications: C18, C45, C51, C63, C87, O13

1. INTRODUCTION

Crude oil products fuel most vehicles in air, on water, and on land all over the world. Fuel derived from Crude oil, such as petrol, kerosene, diesel, heating oil, and so on supply 33% of all the energy consumed by households, businesses, and manufacturers in the world, (BP, 2018). Changes in crude oil price is therefore of significant interest to decision makers especially finance practitioners and commodity market participants. Unfortunately, crude oil price is the most complex and difficult to model because the changes are frequent, nonlinear, irregular and nonstationary. Thus, accurate forecasting of the crude oil price time series is one of the greatest challenges and among the most important issues facing energy researchers and economists towards better decisions at several managerial levels. As a result, achieving reliable and highly accurate forecasting models to answer the uncertainties and complexities of crude oil price is necessary and important to policy makers.

2. LITERATURE REVIEW

Several works have been done on forecasting of crude oil price in recent times. Existing literature review have characterized forecasting techniques into two main groups; quantitative and qualitative methods (Behmiri and Manso, 2013).

The time series of monthly cigarette sales in China have had double trends which include long-term upward trend and seasonal fluctuations trend. The complex time series cannot allow single linear or nonlinear forecasting model capture features of the data, so the results are inaccurate. Autoregressive integrated moving average (ARIMA) and GMDH models were combined to take advantage of their unique strengths in linear and nonlinear modeling respectively. Comparing forecast result with actual data sets confirmed the proposed hybrid model could be an effective way to improve forecasting accuracy

achieved by using either of the models separately (Aiyun et al., 2010).

ARIMA modelling and GMDH neural network, which are quantitative methods were individually used to do a short-term forecast from February 2015 to April 2015 of the prices of four crude oil extracts; Diesel, Kerosene, Petrol, and Liquified Petroleum Gas in India. Forecasted results were compared with the actual prices for the above period. The forecasting accuracy for all the four petroleum products showed promising results which justified the ARIMA and GMDH models forecasting the price of the different petroleum products in India (Khan et al., 2015).

The econometric model ARIMA, a linear model has been used widely for forecasting crude oil price. The ARIMA model is not able to forecast the crude oil price accurately which has non-linear characteristics; the ARIMA model cannot capture all the dynamic properties of the crude oil price hence the ARIMA residual. (Wang et al., 2005; Yu et al., 2008) This paper addressed the problem by the modelling the time series of crude oil using the hybrid ARIMA-phGMDH methodology as shown in the flow chart in Figure 1.

The ARIMA residual which was produced as part of the ARIMA model of crude oil price have been observed to have harmonic properties[1], an example is shown in Figure 2, Residual for ARIMA (0, 2, 2).

phGMDH has the capacity to model harmonic data into a polynomial. (Ivakhnenko and Ivakhnenko, 1995; Nikolaev and Iba, 2003; Onwubolu, 2011; 2014b). Hence the proposed hybrid method is designed to improve the accuracy of ARIMA crude oil price forecasting model. This paper focused on the modelling of a hybrid model and is organized into the introduction, literature review, data, methodology, results and discussion and conclusion.

3. RESEARCH DATA

Daily spot Brent crude oil prices from 2003 to 2017 was used in this paper, (U.S Department of Energy, 2019). On some days, data were not issued due to unknown reasons. These missing data points were filled by the interpolation method (Burden and Faires, 1997a; Moler, 2015a).

4. RESEARCH METHODS

This research paper applied two methodologies in building the hybrid forecasting model; the ARIMA and the phGMDH methods. The motivation for the adaption for the phGMDH is the fact that it can model harmonic time series (Nikolaev and Iba, 2003). The ARIMA methodology on the other hand can model the crude oil price and have harmonic residuals. The hybrid ARIMA-phGMDH methodology is shown in Figure 1.

4.1. The ARIMA Modelling

The ARIMA modeling or Box–Jenkins methodology includes three iterative steps:

1. Model Identification
2. Parameter estimation
3. Diagnostic checking

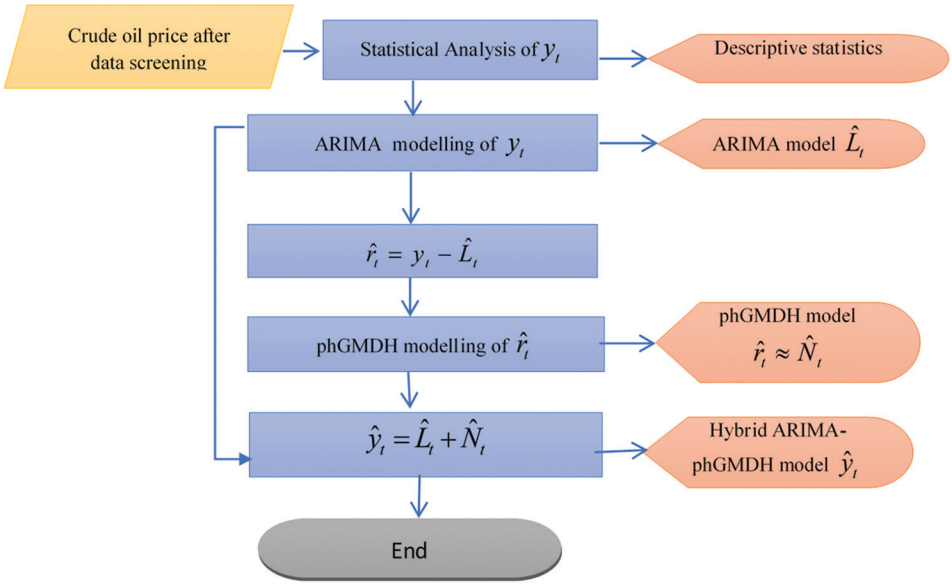
The ARIMA model of order ARIMA (p, d, q) is one of the time series forecasting methods for the non-stationary data series. The ARIMA (p, d, q) can be expressed as in Eq. (1) (BowErman et al., 2005).

$$\phi_p(B)\nabla^d y_t = \delta + \theta_q(B)a_t \quad (1)$$

4.2. phGMDH Modeling on ARIMA Residual

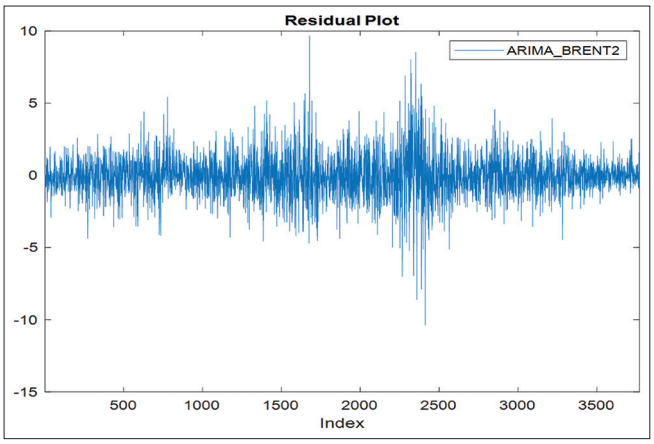
The phGMDH is a variant of the multilayer GMDH network algorithm. It constructs hierarchical layers of bivariate

Figure 1: Autoregressive integrated moving average-phGMDH Methodology flow chart



[1]Refer to (Axler et al., 2013)

Figure 2: Residual for autoregressive integrated moving average (0, 2, 2)



activation polynomials, Eq. (3), in the nodes and variables in the leaves. One or more of the activation polynomials terms is/are harmonic terms of the form as shown in Eq. (2), (Nikolaev and Iba, 2003).

$$\gamma_i = C_i \cos(w_i t - \phi_i) \quad (2)$$

$$p(x_j, x_i) = a_0 + a_1 x_j + a_2 x_i + a_3 x_j x_i \quad (3)$$

The phGMDH modelling process uses simple low-order bivariate activation polynomials, of form Eq. (3). Here the nodes are the hidden units, the leaves are inputs and activation polynomial coefficients are weights. The outcome of the activation polynomial, feeds forward to the parent nodes, creating new partial polynomial models. The algorithm thus churns out high-order multivariate polynomials, Eq. (5), by bringing together simple and tractable activation polynomials allocated in the hidden nodes of the network. For a given data series, Eq. (4), of vectors for all Y_t and x_t that belongs to real number, where α_t are term coefficients (weights), x is an input vector, $x_t = (x_{t-b}, x_{t-b-1}, ..., x_{t-1})$, b, the number of x variables, the objective of the phGMDH algorithm, $F(t,x)$, is to grow a higher order polynomial equation as shown in Eq. (5). Figure 3 demonstrates the phGMDH modelling process.

$$D = [(x_t, y_t)]_{t=1}^{N} \quad (4)$$

$$F(x,t) = a_0 + \sum_i a_i \varphi_i(t,x) + \sum_i \sum_j a_{ij} \varphi_{ij}(t,x) + \sum_i \sum_j \sum_k a_{ijk} \varphi_{ijk}(t,x) + ... \quad (5)$$

The phGMDH modelling is a two-step process;

1) Network initialization
2) Network construction and weight training (Nikolaev and Iba, 2003).

4.3. The Hybrid ARIMA phGMDH Modelling

ARIMA models the linear L_t capabilities in the data y_t whilst phGMDH approximates the non-linear N_t, characteristics of the data using the residual of ARIMA model as input, Thus, $\check{y}_t$, the hybrid approximation of the model is achieved by adding the linear and the nonlinear components as shown in Eq. (6), (Zhang, 2003).

$$\check{y}_t = \check{L}_t + \check{N}_t \quad (6)$$

5. RESULTS AND DISCUSSION

The structure of the results flows as shown in Figure 1: ARIMA-phGMDH Methodology flow chart.

5.1. The ARIMA Modelling

5.1.1. Model identification

The model is identified when the time series is made stationary. Stationary time series is achieved by differencing the time series. The series is made stationary on the second differencing. Figure 4, sample autocorrelation (SAC) and sample partial autocorrelation (SPAC) of ARIMA for second differencing of Brent crude oil price confirms the stationarity. The SAC cut off at lag 2 and the sample SPAC dies down slowly. Hence the model was identified at ARIMA (0, 2, 2)

5.1.2. Parameter estimation

The Brent time series was run on MATLAB econometric modeler (Manjon, 2018), using the model ARIMA (0, 2, 2). The ARIMA (0, 2, 2) model has Eq. (7) as the generic equation.

$$z_t = \delta + (1 - \theta_1 \times B - \theta_2 \times B^2) \times a_t \quad (7)$$

Substituting the parameters from the output of the econometric modeler, using the Brent crude oil price series, produced the tentative Brent ARIMA (0, 2, 2) model, Eq. (8).

$$z_t = a_t + 0.9659 \times a_{t-1} + 0.0341 \times a_{t-2} \quad (8)$$

Figure 2 shows the ARIMA residual of Brent series.

5.1.3. Diagnostic checking

Two tests were performed, the residual SAC (RSAC) and residual SPAC (RSPAC) tests, and the Ljung-Box Q-test, (Ljung and Box, 1978). The RSAC and RSPAC in Figure 5 did not show autocorrelation hence the model was exhaustive (Bowerman, et al., 1993). On the other hand, the Ljung-Box Q-test indicated an inadequate model by rejecting the null hypothesis that is the first m autocorrelations of the residuals of ARIMA (0, 2, 2) are jointly zero. The ARIMA (0, 2, 2) was inadequate, hence the residual, Figure 2 was remodeled. It served as input for the phGMDH modelling.

5.2.The phGMDH Modelling

5.2.1. Step 1: Network initialization

5.2.1.1. Step 1.1 Data organisation

The input data was the ARIMA residual series r_t as shown in Figure 2, with 3768 data points. It was organized into a 628 × 6 matrix of form Eq. (4) such that $x_t = (x_1, x_2, x_3, x_4, x_5)$, with dimension of the independent variables, 628 × 5 and y_t the dependent variable. Figure 6 shows the assigned residual input data for the phGMDH. The assignment of the residual r_t to x_t and y_t was done using Eq. (9).

$$For\, i = 0:627, j = 1:5\, x_{i+1,j} = r_{j+i\times 6}\, y_{i+1} = r_{(i+1)\times 6} \quad (9)$$

5.2.1.2. Step 1.2 spectral analysis

The weighting coefficients α_q, were computed using Eq. (10).

$$\sum_{q=0}^{h-1} \alpha_q (y_{t-q} + y_{t+q}) = y_{t-h} + y_{t+h} \quad (10)$$

There are 208 weighting coefficients, α_q, they are used to characterize and form the non-multiple frequency function Eq. (11) with 209 unique frequencies, w_i.

$$\alpha_0 + \alpha_1 \cos(w_1) + \alpha_2 \cos^2(w_2) + ... + \alpha_h \cos^h(w_h) = 0 \quad (11)$$

The resultant non-multiple frequency function is a 208th degree polynomial. Newton Raphson's method (Burden and Faires, 1997b) or fzero function in MATLAB (Moler, 2015b) is applied to solve for w_i. For, $1 \le i \le h$, $|\cos(w_i)| \le 1$ and $w_i \in$ shows two w_i, non-multiple harmonic frequencies, that are identified to be real numbers within the interval [0, π]. The amplitudes A_i, B_i, resultant amplitude, C_i and phase angles ϕ_i, for each angular frequency w_i is computed (Nikolaev and Iba, 2003). For each of the independent vector variables x_t the intensity plot function is produced using Eq. (12).

$$I(w_i) = \frac{N(A_i^2 + B_i^2)}{4\pi} \quad (12)$$

Dominant harmonics in each independent variable vector, x_t is identified by selecting harmonics with the greatest intensity $I(w_i)$, as shown in Table 1. Harmonics with the highest intensities H_1 and H_3 are chosen to replace x_1 and x_3 respectively to form Eq. (13) the input for the GMDH as shown in Figure 7.

$$D = [(x_t, \gamma_t, y_t)]_{t=1}^{N} \text{ c} \quad (13)$$

5.2.1.3. Step 1.3 Initializing of parameters

Now there are 10 combinations of the input variables for the first layer of the GMDH process. Network parameters width, κ, layer l, lowest error for convergence, ε and activation polynomial are initialized.

5.2.2. Step 2: Network construction and weight training

The input matrix in Figure 7: GMDH input, is fed to the GMDH algorithm in MATLAB (Mohammed, 2014; Onwubolu, 2014a,). The output is the polynomial, Eq. (14).

$$y_t = -0.142458 + 0.031590 \times x_2 - 0.005344 \times H_3 - 0.022953 \times x_2 \times H_3 \quad (14)$$

Interchanging r_i from Eq. (9) for x_i and y_i in Eq. (14) gives the Eq. (15).

$$\check{r}_t = -0.142458 + 0.031590 \times r_{t-4} - 0.005344 \times H_3 - 0.022953 \times r_{t-4} \times H_3 \quad (15)$$

From Table 1

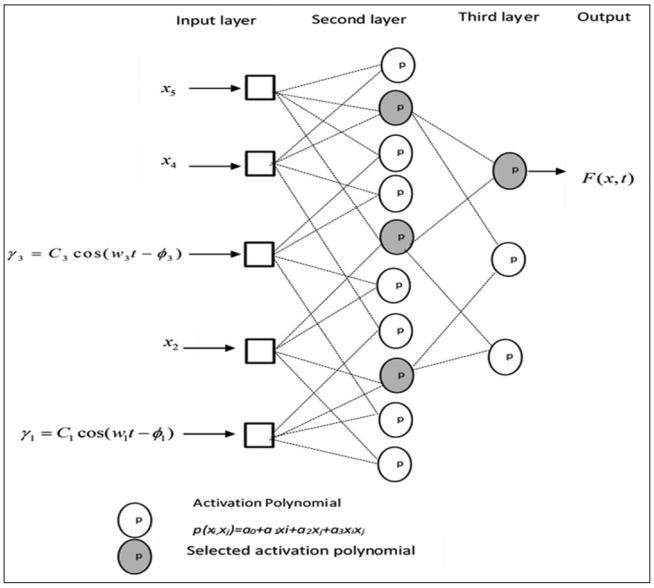
$$H_3 = 6.4809 \times \cos(1.0258 \times t + 1.5262) \quad (16)$$

Substituting Eq. (16) into Eq. (15) gives the Eq. (17)

$$\hat{r}_t = -0.1425 + 0.0316 \times r_{t-4} - 0.0346 \times \cos(1.0258 \times t + 1.5262) - 0.1488 \times r_{t-4} \times \cos(1.0258 \times t + 1.5262) \quad (17)$$

Eq. (17) is the approximation for the residual phGMDH model, $\check{r}_t$.

Figure 3: phGMDH network



5.3. The Hybrid ARIMA phGMDH

The computation at this stage have been made easier as the ARIMA (0, 2, 2) and phGMDH estimate are at a differencing order of 2. Per definitions of Eq. (6), Eq. (8) and Eq. (17)

$$\hat{L} = z_t = a_t + 0.9659 a_{t-1} + 0.034096 a_{t-2} \quad (18)$$

and

$$\hat{N}_t \approx \hat{r}_t = -0.1425 + 0.0316 \times r_{t-4} - (0.0346 + 0.1488 \times r_{t-4}) \times \cos(1.0258 \times t + 1.5262) \quad (19)$$

Table 1: Dominant harmonics in x_t

Harmonics	*I*	*w*	*A*	*B*	*C*	*ϕ*
y	3419.7	0.333	8.1254	1.5514	8.2722	0.18866
H_1	358.76	0.333	−2.0106	−1.771	2.6793	0.72213
H_2	219.85	0.333	−2.0963	0.069304	2.0974	−0.033049
H_3	2099.1	1.0258	0.28882	−6.4745	6.4809	−1.5262
H_4	471.34	0.333	−2.9323	0.91274	3.0711	−0.30176
H_5	2139.2	1.0258	0.040834	−6.5425	6.5426	−1.5646

Figure 4: Sample autocorrelation and sample partial autocorrelation of autoregressive integrated moving average: For second differencing of Brent crude oil price

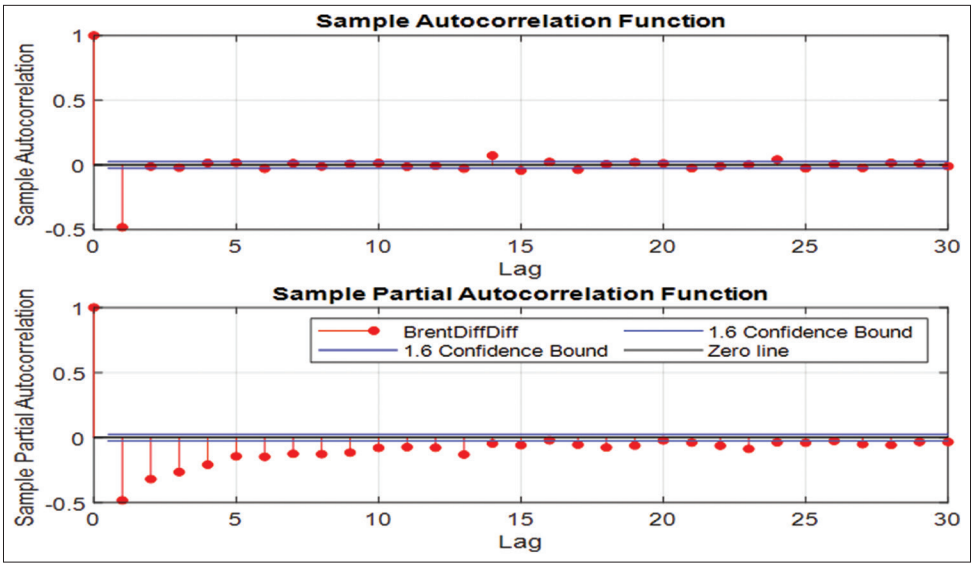
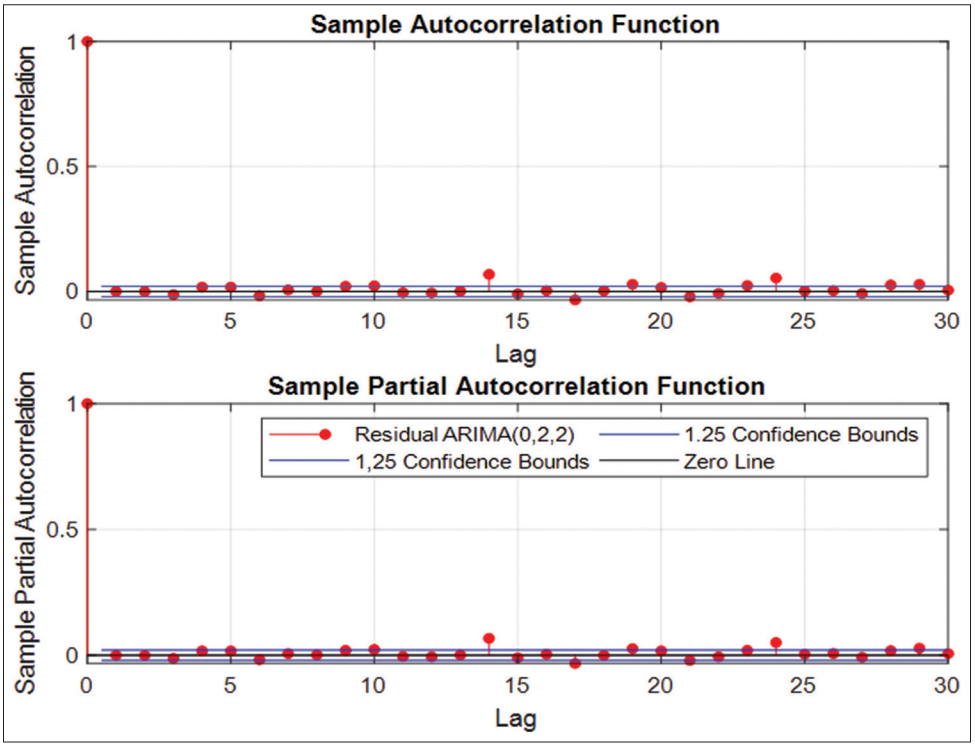


Figure 5: Residual sample autocorrelation and residual sample partial autocorrelation of autoregressive integrated moving average (0, 2, 2)



If Eq. (18) is a linear and Eq. (19) is nonlinear model then the hybrid model, $\hat{y}_t$ is Eq. (20).

$$\hat{y}_t = a_t + 0.9659 \times a_{t-1} + 0.0341 \times a_{t-2} - 0.1425 + 0.0316 \times r_{t-4} - (0.0346 + 0.1488 \times r_{t-4}) \times \cos(1.0258 \times t + 1.5262) \quad (20)$$

6. CONCLUSION

Organisations all over the world have a major duty of care to optimise the use of resources at their disposal to assure their going concern. One major resource that directly or indirectly affect profitability and sustainability of organisations is energy. Crude oil provides at least 33% of all energy needs of most organizations. Forecasting crude oil price is a proactive measure in the process of hedging against crude oil price risk, a major influence on most organisations' sustainability.

This study is an attempt to close the gap in forecasting the crude oil price using ARIMA methodology by building a hybrid ARIMA-phGMDH model to forecast Brent crude oil price using historical data. This paper demonstrates that it is feasible to build a hybrid forecasting model using the ARIMA and phGMDH models. The developed model for Brent crude oil as shown in Eq. (20) can be simulated to forecast future crude oil price for short term to medium

Figure 6: Residual input for the phGMDH

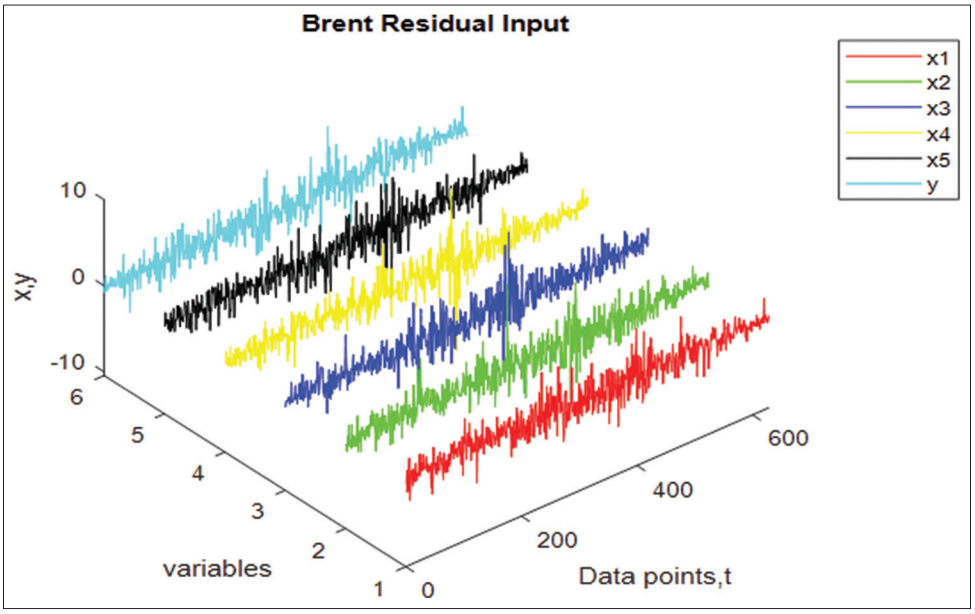
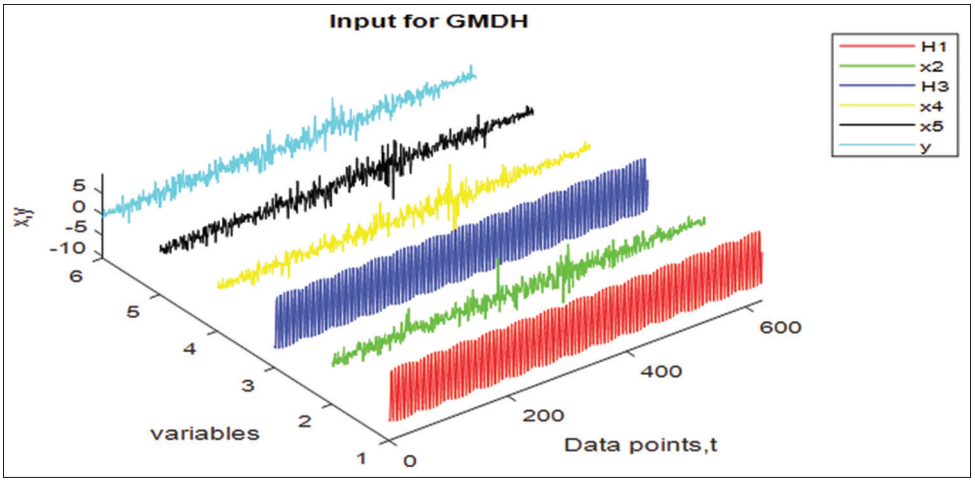


Figure 7: GMDH input



term periods. This concept can be extended to the other crude oil markets in the oil and gas industry. The data can be updated to accommodate change in current data. Literature have confirmed an above 95% efficacy of the individual model components, the ARIMA and phGMDH models. Future work will concentrate on evaluating the ARIMA-phGMDH model to assess its forecasting efficacy.

7. ACKNOWLEDGMENTS

This research has been conducted in Universiti Teknologi Petronas, where the first author was supported by the School's Graduate Assistantship Scheme. Am also grateful to the Faculty of Fundamental and Applied Science Department for their valuable criticisms and contribution in this research.

REFERENCES

Aiyun, Z., Weimin, L., Fanggeng, Z. (2010), Double Trends Time Series Forecasting using a Combined ARIMA and GMDH Model. In 2010 Chinese Control and Decision Conference. p1820-1824.

Axler, S., Bourdon, P., Wade, R. (2001), Basic properties of harmonic functions. In: Harmonic Function Theory. 2nd ed. New York: Springer-Verlag Inc. p1-25. Available from: http://www.images.wikia.com/nccmn/ro/images/8/80/Harmonic_Function_Theory.pdf.

Behmiri, N.B., Manso, J.R.P. (2013), Crude Oil Price Forecasting Techniques: A Comprehensive Review of Literature. SSRN. Available from: https://www.doi.org/10.2139/ssrn.2275428.

Bowerman, B.L., O'Connell, R.T., Koehler, A.B. (1993), Estimation, diagnostic checking and forecasting for nonseasonal box-jeckins model. In: Forecasting and Times Series: An Applied Approach. 3rd ed. USA: Duxbury Thomson Learning. p488-520.

Bowerman, B.L., O'Connell, R.T., Koehler, A.B. (2005), Box-jenkins seasonal modeling. In: Day, A., Brayton, K., editors. Forecasting and Time Series: An Applied Approach. 4th ed. USA: Brooks/Cole Thomson. p490-513.

BP, (British Petroleum). (2018), BP Statistical Review of World Energy June 2018. 67th ed., Vol. 2018. Available from: https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html.

Burden, R.L., Faires, J.D. (1997a), Interpolation and polynomial

approximation. In: Ostedt, G., editor. Numerical Analysis. 6th ed. USA: Brooks/Cole Publishing Company. p104-165.

Burden, R.L., Faires, J.D. (1997b), Numerical solutions of nonlinear systems of equations. In: Ostewdt, G., editor. Numerical Analysis. 6th ed. USA: Brooks/Cole Publishing Company. p588-621.

Ivakhnenko, A.G., Ivakhnenko, G.A. (1995), The review of problems solvable by algorithms of the group method of data handling (GMDH). Pattern Recognition and Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii, 5, 527-535.

Khan, M.S., Saikia, A., Tripathy, S. (2015), Prediction of petroleum price in India. International Journal of Research in Mechanical Engineering and Technology, 5(2), 11-16.

Ljung, G.M., Box, G.E.P. (1978), On a measure of lack of fit in time series models. Biometrika, 65(2), 297-303.

Manjon, J. (2018), Econometric Modeler. Massachusetts, United States: MathWorks Inc. Available from: https://www.uk.mathworks.com/products/econometrics.html?s_tid=AO_PR_info.

Mohammed, M.A.A. (2014), GMDH multi-layered algorithm in MATLAB. In: GMDH-Methodology and Implementation in MATLAB. London: Imperial College Press. p75-124.

Moler, C. (2015a), Numerical Computing with MATLAB. Available from: https://www.mathworks.com/content/dam/mathworks/mathworks-dot-com/moler/interp.pdf. [Last retrieved on 2019 Mar 20].

Moler, C. (2015b), Numerical Computing with MATLAB. Available from: https://www.mathworks.com/content/dam/mathworks/mathworks-dot-com/moler/zeros.pdf. [Last retrieved on 2019 Mar 21].

Nikolaev, N.Y., Iba, H. (2003), Polynomial harmonic GMDH learning networks for time series modeling. Neural Networks, 16(10), 1527-1540.

Onwubolu, G.C. (2011), GMDH harmonic algorithm. In: GMDH-Methodology and Implementation in C. London: Imperial College Press. p93-106.

Onwubolu, G.C. (2014a), GMDH multi-layered algorithm. In: GMDH-Methodology and Implementation in MATLAB. London: Imperial College Press. p27-74.

Onwubolu, G.C. (2014b), Introduction. In: GMDH-Methodology and Implementation in MATLAB. London: Imperial College Press. p3-26.

U.S Department of Energy, Brent Crude Spot Price, Energy Information Administration. (2019), Available from: https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=RBRTE&f=D. [Last accessed on 2019 Jan 31].

Wang, S., Yu, L., Lai, K.K. (2005), In: Xu, W., Chen, Z., editors. A Novel Hybrid AI System Framework for Crude Oil Price Forecasting. Berlin, Heidelberg LB: Springer. p233-242.

Yu, L., Wang, S., Lai, K.K. (2008), Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics, 30(5), 2623-2635.

Zhang, G.P. (2003), Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.