Jurnal Teknologi

CRUDE OIL PRICE FORECASTING BY CEEMDAN BASED HYBRID MODEL OF ARIMA AND KALMAN FILTER

Muhammad Aamir[a]*, Ani Shabri[a], Muhammad Ishaq[b]

[a]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
[b]Department of Mathematics, School of Natural Sciences, National University of Sciences and Technology, Islamabad, Pakistan

Graphical Abstract
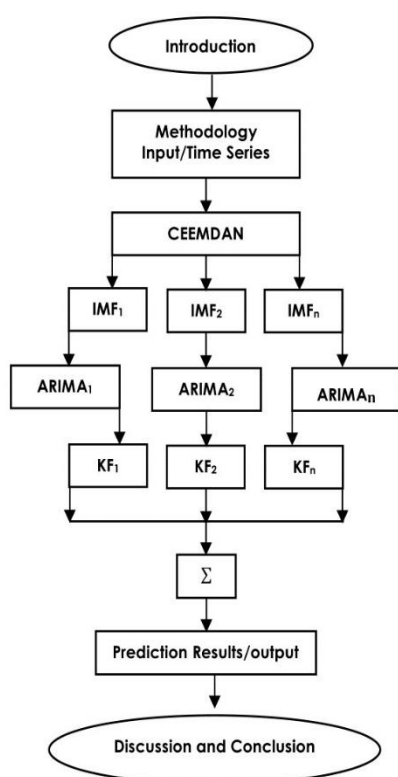
Abstract

This paper used complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) based hybrid model for the forecasting of world crude oil prices. For this purpose, the crude oil prices original time series are decomposed into sub small finite series called intrinsic mode functions (IMFs). Then ARIMA model was applied to each extracted IMF to estimate the parameters. Next, using these estimated parameters of each ARIMA model, the Kalman Filter was run for each IMF, so that these extracted IMFs can be predicted more accurately. Finally, all IMFs are combined to get the result. For testing and verification of the proposed method, two crude oil prices were used as a sample i.e. Brent and WTI (West Texas Intermediate) crude oil monthly prices series. The D-statistic values of the proposed model were 93.33% for Brent and 89.29% for WTI which reveals the importance of the CEEMDAN based hybrid model.

Keywords: ARIMA, CEEMDAN, Crude Oil, EMD, Kalman Filter

Abstrak

Kajian ini merupakan kombinasi penuh penguraian mod emperik dengan penyesuaian bunyi atau ensemble empirical mode decomposition with adaptive noise (CEEMDAN) berdasarkan kepada kombinasi model ramalan harga minyak mentah dunia. Untuk itu, siri harga minyak mentah asal di bahagikan kepada subkecil terhingga yang diberi nama intrinsic mode functions (IMFs). Kemudian, model ARIMA digunakan oleh setiap IMF untuk penganggaran parameter. Dengan nilai penganggaran parameter model ARIMA, Kalman Filter digunakan oleh setiap IMFs bagi mendapatkan jangkaan bacaan yang tepat. Untuk pengujan dan pengesahan, dua harga minyak mentah digunakan. Contohnya dengan peranan siri harga minyak mentah dari Brent dan WTI (West Texas Intermediate). Nilai D-statistic yag dicadangkan pada model Brent adalah 93.33% dan untuk WTI adalah 89.29% dimana ianya menunjukkan kepentigan CEEMDAN berasaskan model hibrid.

Kata kunci: ARIMA, CEEMDAN, EMD, Kalman Filter, Minyak Mentah

1.0 INTRODUCTION

Crude oil is a very important commodity in the world because of his unique nature, it affects every individual's life in many ways. Everybody knows that oil is a non-transferable commodity but its use increases day by day irrespective of the fact that it is a non-renewable commodity. As compared to other commodities oil needs more attention due to his high demand in every field. Due to the changing aspects of oil prices, it is a challenge for econometrician, mathematician, and statistician to design a better strategy for the industrialists who consume or supply oil can take up-to-date and more accurate decisions. Up to date decisions is not only important for industrialists but also more important for investors, suppliers and government agencies for planning their activities within the available resources. Crude oil is treated as a special commodity among different world players such as oil companies, oil producing nations, oil importing nations, speculators and individual refineries. The crude oil price fluctuations involve so many factors like supply, inventory, consumption and demand but irrespective of these factors the oil price also is influenced by irregulars and unpredictable elements which are random in nature. These characteristics of the crude oil price make a large fluctuation in the market Watkins and Plourde[1]. Due to the irregular and stochastic nature of oil prices, it is 'a very complex and challenging task for researchers to develop appropriate models for forecasting the crude oil prices. The complex and compound nature of crude oil price makes this area widely opened for researchers to develop many different procedures for forecasting the crude oil prices in a good manner. As far as the world economy is concerned crude oil plays an increasingly very important role in the world energy demands which is met by two third from the crude oil Alvarez-Ramirez *et al.* [2]. Verleger [3] said that crude oil is the most active world largest traded commodity which is accounted for more than ten percent of the world total trade. Like other commodities, the price of crude oil is also measured from the demand and supply Hagen[4] and Stevens[5]. Additionally, the whole aggregate economic activity can be disturbed by a sharp movement of crude oil price which can fluctuate the nation's economy significantly; further, the impacts can be reflected in two ways on nation's economy. Firstly, the oil importing countries economic growth is harmfully influenced and increase inflation by a sharp increase in crude oil prices. Secondly, the oil exporting countries face a serious budgetary shortfall problems by a small fall in crude oil prices Abosedra and Baghestani[6]. The topic of forecasting crude oil price is very important although it is tremendously hard one, due to its high volatility and inherent difficulties Wang[7].

In the last two decades, typical statistical tools and econometric methods were used for the forecasting of crude oil prices, such as GARCH models, ARIMA models, linear regression, naïve random walk, ECM (error correction model) and VAR (vector autoregressive) models. For example, a refine econometric model was used by Huntington[8] to predict the oil prices in 1980's. A probabilistic approach was used by Abramson and Finizza[9] for forecasting the crude oil prices. Barone-Adesi *et al.* [10] proposed a semi-parametric approach for prediction of crude oil prices while Morana [11] also proposed a semi-parametric statistical technique for forecasting the crude oil price for a short term using the GARCH properties. For prediction of WTI prices, the co-integration procedure was used by Gülen[12]. To predict the U.S. oil price VAR model was used by Mirmirani and Li[13] while an investigation of crude oil and their products prices ECM model was used by Lanza *et al.* [14]. Using inventory levels of OECD petroleum, comparative inventories, and low and high inventories variables for the WTI prices the simple econometric models was used by Ye *et al.* [15, 16, 17]. Dées *et al.* [18] also, used cause and effect model for the forecasting of crude oil prices. Other researchers such as Liu[19], Chinn *et al.* [20], Agnolucci[21] and Ahmad[22] used a well-known Box-Jenkin's methodology for forecasting the crude oil prices. Furthermore, some studied have been found which used the GARCH type of models for forecasting the Crude oil prices namely Sadorsky[23], Hou and Suardi[24], Ahmed and Shabri[25]. Aamir and Shabri [26] used the hybrid model of ARIMA GARCH for Pakistan crude oil prices. For best predictions, different approaches are used and when there are competing models the model is selected on the basis of past accuracy, but the problem is still there when the differences are significant. Diebold and Mariano [27] discussed that decisions are more critical in the situations, on determining the model selection for which predictions they take their decisions.

The other approach used for univariate methods is called state space modelling which is also called dynamic linear modelling. The use of these models makes possible help in common mathematical problems in the stage of model development Hyndman *et al.* [28]. A lot of statistical tools can be accessed by putting the model in state space form, is one of the main advantages of this approach. The Kalman Filter technique can be used for obtaining the optimal forecast and unobserved components best estimates are achieved by the Kalman Filter smoothing Harvey[29]. Kalman [30] introduced the Kalman Filter for the first time and due to this technique, the modern work of state space modelling was started and at its earlier stages, it was proved that this approach is quite helpful in the field of space technology and engineering. It was stated that there is no general method available for implementation of Kalman Filter in the field of statistical forecasting Morrison and Pike[31]. However, the researchers started their work for implementation of Kalman Filter technique in statistics. The importance of Kalman Filter in the field of econometrics and statistics, for example, time series forecasting was discussed by

Rosenberg[32], Engle[33], Harvey and Phillips[34]. Ravichandran and Prajneshu [35] emphasized the importance of this procedure. Nikolaisen Sävås [36] stated that the Kalman Filter technique produced the best forecast for four different countries inflation data as compared to the SARIMA models. Recently, Aamir and Shabri[37]found that the hybrid model of ARIMA Kalman Filter produced better forecast than the hybrid model of ARIMA GARCH and single ARIMA model for Pakistan crude oil prices.

Whenever an ARIMA model is used for a financial time series, one can remember that these series are non-linear and non-stationary in nature and the ignorance of these properties will produce worse predictions. The divide and conquer principle Yu *et al.* [38] for reconstructing of financial time series are very important for forecast models. To solve these limitations of financial time series the concept of hybrid models was introduced and widely used these days. Empirical Mode Decomposition (EMD) is one of the procedures which divide the financial time series into sub-simple series for which the forecasting will be quite simple. EMD decomposed the non-stationarity and non-linearity behaviour of the financial time series into independent sub time series Tang *et al.* [39]. The EMD can also expose the trend and the hidden patterns of the time series which can be easily handled for forecasting models in different time series applications An *et al.* [40] and Guo *et al.* [41]. For example, Yu and Wang [38] proposed a hybrid model of EMD-FNN-ALNN for forecasting of Brent and WTI crude oil prices, whereas Guo and Zhao[41] used a hybrid model of EMD-FFNN for forecasting the wind speed time series. Despite these facts, EMD has still a limitation of "mode mixing problem" and it arises when clear spectral separation of modes is not achieved Huang and Wu[42] and Schlotthauer *et al.* [43]. In order to avoid this difficulty, Wu and Huang [44] propose a new technique called EEMD (ensemble empirical mode decomposition), its usefulness has proven for some geophysical data sets, for example, Wu *et al.* [45]. But EEMD has still some problems of mode mixing and residual noise in some real data sets. These drawbacks encouraged Torres *et al.* [46] to suggest an alternative method of EEMD which is called CEEMDAN (complete ensemble empirical mode decomposition with adaptive noise), which produces a more clear spectral separation of modes and exact reconstruction of the actual signal. Due to these extra strengths, CEEMDAN became a useful tool for extracting some meaningful information from financial time series, in situations when the aforesaid complications are faced. Taking the above notes regarding the CEEMDAN Antico *et al.* [47] suggested that it performed well as compared to EMD and EEMD and extract more clear modes. CEEMDAN technique has not been used till now for such type of data set up to the best of our knowledge. This paper focused on forecasting the two well-known crude oil prices time series. In this paper, new technique CEEMDAN is used for the decomposition of non-stationary data into some meaningful modes and trends.

2.0 METHODOLOGY

In this section, the whole methodologies regarding all models are presented. First, the ARIMA technique was reviewed briefly. The second is state space models which are used by Kalman Filter technique. The third is the EMD mode decomposition technique and the last is CEEMDAN mode decomposition procedure. After these techniques, the next is evaluation criteria's namely MAE, RMSE, MAPE and $D_{statistic}$ be presented briefly.

2.1 ARIMA

The models which require that the time series should be stationary is the ARIMA models. Stationarity implies that the mean, variance and auto-correlation structure of the time series remains constant over the time because the data is smooth and having no seasonality and trend factors. The successive differences have been taken whenever the given time series is not stationary. ARIMA models are preferably used for the time series measurement, which has some serial dependence. The Interdependencies in time series (x) is measured by the auto-regressive (AR) terms while the dependence on preceding error terms measured by moving average (MA) terms Montgomery *et al.* [48]. An ARMA (Auto-regressive Moving Average) ARIMA model of order (k, m) for a univariate series has the following form:

$$x_t = x_0 + \alpha_1 x_{t-1} + \cdots + \alpha_k x_{t-k} + \epsilon_t + \beta_1 \epsilon_{t-1} + \cdots + \beta_r \epsilon_{t-m}$$

For selection of the appropriate order of polynomials, the sample auto-correlation and partial-autocorrelation functions were used and compared with hypothetical bounds of Box *et al.* [49].

2.2 Kalman Filter

Before introducing the Kalman Filter technique, some introduction of the state space modelling is necessary because it uses the state space model terminology. The state space model works in two steps. Firstly, a state vector is created for capturing the substantial components of the series and added till the end. The state vector which is also called the smallest vector of the system plays a very important role in the state space modelling which summarizes the historical behaviour of overall system Brocklebank and Dickey[50]. The two linear equations explain the smoothing of the series in state space modelling. The first equation called observation equation comprised of the present observation and unobserved state and the second equation which is called state equation determines the progress over time in states and update the state vector continuously. Basically, the state equation is a vector containing the unobserved factors of the series, such as seasonality and trend. The state space models are presented in the following form:

$$x_t = T'\gamma_t + \varepsilon_t$$
$$\gamma_t = H'\gamma_{t-1} + F\omega_t$$

Where x_t is the observed vector of variables, γ_t is the vector of unobserved variables, F, H and T are parametric matrices, ε_t and ω_t are the white noise terms with variance-covariance matrices Q and R respectively Hamilton[51].

Kalman Filter is the best linear estimator and infers the parameters from inaccurate and uncertain observation. The Kalman Filter updates the system after receiving each new observation and this feature differentiates the Kalman Filter from other techniques. Hence due to updating after each observation the mean square error of the state vector minimizes. The best estimates are found for the state vector by filtering the noise term and update the state vector. Once the state vector is updated it revises gradually the mean moments of the state vector distribution as well as predicts the unobserved time series values. The Kalman Filter is a procedure for elucidating the linear state space models. Thus, the updating and predicting the system of equations is called Kalman Filter. Kalman Filter retains the state vector estimates and covariance matrix of the state errors Harvey[29]. The innovative portion of the time series x_t, which is not described from the past behaviour of the series called innovations and equivalent to the residuals of Kalman Filter. The Kalman filter predicts recursively the states estimates of the series x_t based on past information with the variance of the prediction error. ε_t is the novelty at time t, which is the new information in series x_t and supposed to be not predicted from the previous information are referred to the one-step ahead forecasting error Saini and Mittal[52]. The simple Kalman Filter recursions are described in three steps below:
(a) Step 1. At time $t = 0$, guess the starting values for state vector $\gamma_{0/0}$ and for variance $P_{0/0}$.
(b) Step 2. Prediction. Use the estimated value of $\gamma_{1/0}$ at $t = 1$ make a best estimate for $x_{1/0}$.
(c) Step 3. Updating. At time $t = 1$ compute the prediction error by using estimated value of x.
The prediction error ($w_{1/0} = x_1 - x_{1/0}$) retains the new information that will be used to refine the guess about the state estimator γ. The estimated value of the state is $\gamma_{1/1} = \gamma_{1/0} + K_t w_{1/0}$, where K_t is Kalman gain (this is the weight allocate to new information) and $\gamma_{1/1}$ is condition on time $t = 1$. The Kalman Filter overall systems are as follows:

At time $t = 0$ starting values	$\gamma_{0/0}, P_{0/0}$
At time $t = 1$, State vector prediction	$\gamma_{t/t-1} = H\gamma_{t-1/t-1}$, $P_{t/t-1} = HP_{t-1/t-1}H' + Q$
Prediction error computation	$w_{t/t-1} = x_t - x_{t/t-1}$ $= x_t - F_t\gamma_{t/t-1}$ $f_{t/t-1} = F_t P_{t/t-1}F' + Z$
Updating States	$K_t = P_{t/t-1}F'f_{t/t-1}^{-1}$

$$\gamma_{t/t} = \gamma_{t/t-1} + K_t w_{t/t-1}$$
$$P_{t/t} = P_{t/t-1} - K_t F_t P_{t/t-1}$$

2.3 EMD

The EMD is a nonlinear signal transformation procedure introduced by Huang et al. [53]. This method is used to molder a non-stationary and nonlinear time series data into IMFs (intrinsic mode functions) with single intrinsic time measure properties. Huang and Shen [53] said that the two conditions must satisfy by each IMF. The first condition is that the number of extreme and zero crossings values must be equal or not differ by at most one value, and the second condition is that the average value of the envelop is zero at any point assembled by local minima and maxima. The decomposition procedure of a time series data are as follows:
(a) Detect all local minima and maxima of the series x_t.
(b) Compute the lower envelope $x_t(l)$ and upper envelop $x_t(u)$ for the series x_t.
(c) Use the lower and upper envelop to obtain the first mean time series $m_t(1)$, i.e. $m_t(1) = (x_t(l) + x_t(u))/2$.
(d) To get the first IMF $imf_t(1)$, find the difference between the series x_t and the mean series $m_t(1)$. i.e. $imf_t(1) = x_t - m_t(1)$. Furthermore, the two conditions will be checked for $imf_t(1)$, if it does not satisfy both conditions, then from step (a) to (c) of decomposition process will be repeated to obtain first IMF.
(e) Repeat the above steps to get all IMFs, until the final residue r_t, which is a monotonic function proposed for the discontinuing of the decomposition process Huang et al. [54].

To get back the original time series x_t from the IMFs and residue simply add these components using this equation. $x_t = \sum_i^n imf_t(i) + r_t$, where n is the total number of IMFs.

2.4 CEEMDAN

Before starting with the CEEMDAN methodology, a brief introduction of EEMD (Ensemble Empirical Mode Decomposition) method is necessary. The EEMD procedure is an improved form of EMD, the ultimate modes are demarcated as mean of the IMFs attained through EMD over an ensemble of trails y_j ($j = 1,2,...,n$), obtained by summing different white-noise realizations ϵ_j, to the original time series x_t. More specifically, Huang andShen[54] elaborate the EEMD procedure are as follows:
(a) Generate $y_j = x_t + \theta\epsilon_j$, where $\epsilon_j(j = 1,2,...,n)$ has a white noise Gaussian process with zero mean and unit variance, and θ is added noise term.
(b) Decompose every y_j ($j = 1,2,...,n$) by using EMD technique into modes IMF_b^j, where $b = 1,2,...,B$ identifies the number of mode.

(c) Assign $\overrightarrow{IMF}_b$ as b^{th} mode of x_t and accomplish the mean of corresponding IMFs. i.e. $\overrightarrow{IMF}_b = \frac{1}{n}\sum_{j=1}^{n} IMF_b^j$

Also, noted that $\overrightarrow{IMF}_b$ are not required to satisfy the IMF properties. In EEMD each y_j is independently decomposed for all realizations. Hence, for each realization different residues are computed $red_b^j = red_{b-1}^j - IMF_b^j$ ($red_0^j = y_j$).

In CEEMDAN Torres andColominas[46] method, the ultimate b^{th} decomposition mode is represented as $\overset{\frown}{IMF}_b$ and to compute the complete decomposition, the first residue is calculated by the proposed method of Colominas *et al.* [55] and Antico and Schlotthauer[47] are as follows: $red_1 = x_t - \overset{\frown}{IMF}_1$ where $\overset{\frown}{IMF}_1$ calculated in the same way like in EEMD. The ensemble red_1 plus different realizations modes are generated by the EMD method with the given noise for each first element. After averaging these first EMD modes to get the second $\overset{\frown}{IMF}_2$, the second residue is calculated as like the first that is: $red_2 = x_t - \overset{\frown}{IMF}_2$ and this procedure continues until the stopping criterion is reached for the rest of modes. To describe the whole CEEMDAN procedure, define an operator $E_b(.)$. The signal giving by $E_b(.)$ produces b^{th} mode obtained through EMD procedure. Let ϵ_j ($j = 1,2,…,n$) be a white noise realization with zero mean and unit variance of Gaussian process. If the targeted data is x_t, the procedure describe by Colominas andSchlotthauer[55] are as follows:

(a) By EMD method decompose n realizations $x_t + \theta_0\epsilon_j$ for the first mode to obtain and calculate $\overset{\frown}{IMF}_1 = \frac{1}{n}\sum_{j=1}^{n} IMF_1^j = \overline{\overline{IMF}}_1$

(b) Compute the first residue for $b = 1$, using $red_1 = x_t - \overset{\frown}{IMF}_1$

(c) For the first each EMD mode the realizations $red_1 + \theta_1 E_1(\epsilon_j), j = 1,2,…,n$ are decomposed and the second mode has follow the given procedure: $\overset{\frown}{IMF}_2 = \frac{1}{n}\sum_{j=1}^{n} E_1\left(red_1 + \theta_1 E_1(\epsilon_j)\right)$.

(d) For $b = 2,3,…,B$ compute the bth residue are as follows: $red_b = red_{b-1} - \overset{\frown}{IMF}_b$.

(e) Continue the process to obtained the first EMD mode of the decompose realizations $red_b + \theta_b E_b(\epsilon_j), j = 1,2,…,n$, the next mode $(b + 1)$ is defined as: $\overset{\frown}{IMF}_{b+1} = \frac{1}{n}\sum_{j=1}^{n} E_1\left(red_b + \theta_b E_b(\epsilon_j)\right)$.

(f) Continue from the step (d) for the next b.

Repeat step *(d)* to *(f)* until the last residue is computed which are no more decomposed. The last residue will satisfy; $red = x_t - \sum_{b=1}^{B} \overset{\frown}{IMF}_b$, where B be the total number of modes, and the original signal from CEEMDAN decomposition can be obtained as: $x_t = \sum_{b=1}^{B} \overset{\frown}{IMF}_b + red$.

In CEEMDAN algorithm the added level of noise is represented by θ_b, whereas different set of values of θ_b can be used for every mode. But in the present study a fixed value of θ_b is used. Colominas and Schlotthauer [55] discussed that the number of realizations and the level of added noise can vary from application to application. In this study the noise level 0.21 is used for the Brent time series while 0.19 used for WTI time series. The applications and implementations of CEEMDAN are available on the following link used in this study http://www.bioingenieria.edu.ar/grupos/ldnlys/.

2.5 Evaluation Criteria's

To measure the forecasting accuracy, different evaluation criterions are used for the level of prediction and directional forecasting. For the level of prediction evaluation, three methods are used, the first one is mean absolute error (MAE), the second is the root mean square error (RMSE) and third is mean absolute percentage error (MAPE) typically these methods are defined as:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |\hat{x}_t - x_t| \,,$$

$$RMSE = \frac{1}{n}\sqrt{\sum_{t=1}^{n}(\hat{x}_t - x_t)^2}$$

and $MAPE = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{\hat{x}_t - x_t}{x_t}\right| \times 100$

Where $\hat{x}_t$ represents the forecasted value while x_t used for the original value and n used for the total number of predictions. In this study for the percentage comparison of MAE, RMSE and MAPE the percentage relative efficiency (PRE) is also calculated for the different models as compared to the basic ARIMA model the formulas used for the PRE are as follows:
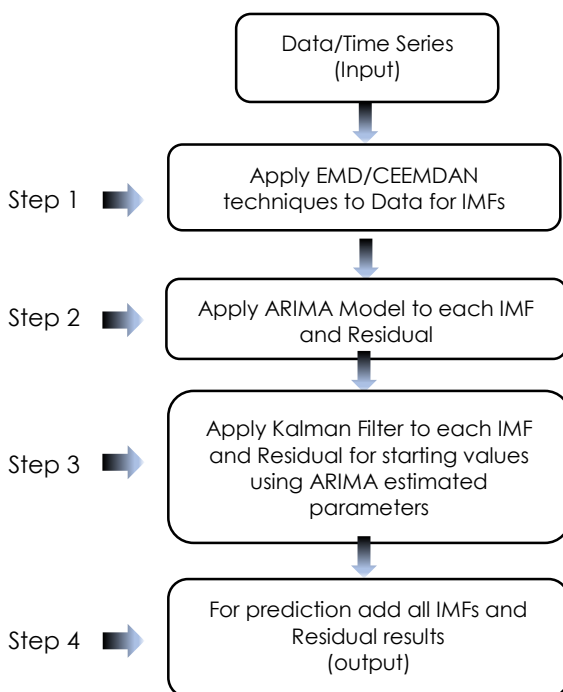

Figure 1 Complete Process of the EMD/ CEEMDAN, ARIMA and Kalman Filter Technique

$$PRE(MAE) = \frac{MAE(ARIMA\ model)}{MAE((\vartheta)model)} \times 100$$

$$PRE(RMSE) = \frac{RMSE(ARIMA\ model)}{RMSE((\vartheta)model)} \times 100$$

$$PRE(MAPE) = \frac{MAPE(ARIMA\ model)}{MAPE((\vartheta)model)} \times 100$$

where ϑ = ARIMA Kalman Filter, EMD ARIMA, EMD ARIMA Kalman Filter, CEEMDAN ARIMA and CEEMDAN ARIMA Kalman Filter

Obviously, predicting models accuracy is one of the best important measures, but the researcher still wants more improvement on the decisions for this purpose the directional predictions are used. For investors, the directional predictions are more important aiming to support more accurate decisions for the purpose of making more money. So, for the crude oil price predictions usually depends on the directions of the original price(x_t) and the forecasted price($\hat{x}_t$). For the directional purpose the $D_{statistic}$ is used Yu and Wang[38], Yu *et al.* [56, 57], which can be defined as;

$$D_{statistic} = \frac{1}{n}\sum_{t=1}^{n} \mu_t \times 100$$

Where $\mu_{t=}1, if\ (x_{t+1} - x_t)(\hat{x}_{t+1} - x_t) \geq 0$ other wise $\mu_t = 0$. Regarding the forecasting competence of the proposed CEEMDAN ARIMA Kalman Filter ensemble methodology with the other models, the ARIMA model is used as a bench mark model. The complete flow chart of the process to be used in this study is presented in Figure 1. In this section the methodology used for the new suggested median estimator and previous estimators suggested by various authors will be discuss in details.

3.0 RESULTS AND DISCUSSION

3.1 Data information

As we know that there are a lot of crude oil price series. In this study, the two-main benchmark crude oil price series are used as a sample. The first series is Brent crude oil prices and the second series is WTI (West Texas Intermediate) crude oil prices. The Brent and WTI crude oil prices are the most common and important benchmark prices and widely used as the basis of so many other crude oil prices formulas. The data used in this study is the monthly data sets for both crude oil prices, and easily obtainable from the DataStream (Thomson Reuters).

For Brent crude oil price, we take the monthly data from January, 1982 to November, 2015 a total of 407 observations. The data from January, 1982 to February, 2011 is used as a training dataset consisting

of 350 observations and the rest 57 observations starting from March, 2011 is used as a testing data set. For WTI crude oil price, the data covers the period from January, 1982 to November, 2015 with a total of 407 observations. Similarly, for WTI series the data from January, 1982 to Feb, 2011 is used as a training data set while the last 57 observations are used as a testing data set starting from March, 2011. Both the testing periods is used to evaluate the performance of the prediction, based on some evaluation criteria's. In this study, only the one step ahead forecasting is carried out.

3.2 Analysis

In this study, R software is used for all the statistical computing which is freely available and supported by the R Foundation for Statistical Computing. For ARIMA model selection and forecasting the FitAR and forecast packages are used. For empirical mode decomposition and for complete ensemble empirical mode decomposition with adaptive noise the Rlibeemd R packages is used. While for Kalman Filter the dlm package of R is used proposed by Petris[58] for dynamic linear models. In this study, the ARIMA model is used as a benchmark for other models. For ARIMA modelling the well-known Box-Jenkins Box and Jenkins[49] methodology is used and for selecting the best model AIC criteria Akaike[59] is used. Thus, for Brent crude oil the best-selected model for training period is ARIMA (4, 1, 3) and for WTI the best ARIMA model of order (2, 1, 3) is chosen and their estimated coefficients are presented in Table 1.

Table 1 The ARIMA model estimation results for the monthly prices of Brent series

Coefficient	Estimated Value	Std.error	t-value	P-value
Brent ARIMA (4,1,3)				
AR (1)	0.3794	0.1753	2.16	0.0304
AR (2)	0.9047	0.0848	10.7	0.0000
AR (3)	-0.3689	0.1553	-2.37	0.0175
AR (4)	-0.1973	0.0744	-2.65	0.0080
MA (1)	-0.1676	0.1761	-0.92	0.3553
MA (2)	-0.8671	0.0563	-15.4	0.0000
MA (3)	0.2267	0.1665	1.36	0.1733
WTI ARIMA (2,1,3)				
AR (1)	1.5292	0.0957	15.9	0.0000
AR (2)	-0.6512	0.0899	-7.2	0.0000
MA (1)	-1.5093	0.1008	-14.9	0.0000
MA (2)	0.8122	0.1190	6.8	0.0000
MA (3)	0.2312	0.0606	-3.8	0.0001

The ARIMA-Kalman Filter model used the estimated values of the ARIMA model presented in Table 1 as starting values of the Kalman Filter recursion. The EMD-ARIMA model uses the ARIMA model to predict each IMF and residue extracted by EMD and after that simply combine all predicted IMFs and residue to get the result. In EMD-ARIMA-Kalman Filter model first ARIMA model apply to the extracted IMFs and residue and obtain the parameters estimate for each ARIMA model after that using the estimated value of ARIMA as starting values for the Kalman recursion and predict all IMFs including residual at the end simply combine the predicted values of all IMFs to get the result. In

as discussed in the case of EMD, here the only difference is that IMFs and residue are extracted by the CEEMDAN technique. While in the case of CEEMDAN-ARIMA-Kalman Filter model, the only difference is the extraction of IMFs and residual as compared to EMD-ARIMA-Kalman Filter model here the IMFs are extracted by the CEEMDAN technique. Thus, ARIMA, EMD-ARIMA ensemble, EMD-ARIMA-Kalman Filter ensemble, CEEMDAN-ARIMA ensemble and CEEMDAN-ARIMA-Kalman Filter ensemble models are all used for comparison to predict the crude oil prices for the two benchmark Brent and WTI respectively.
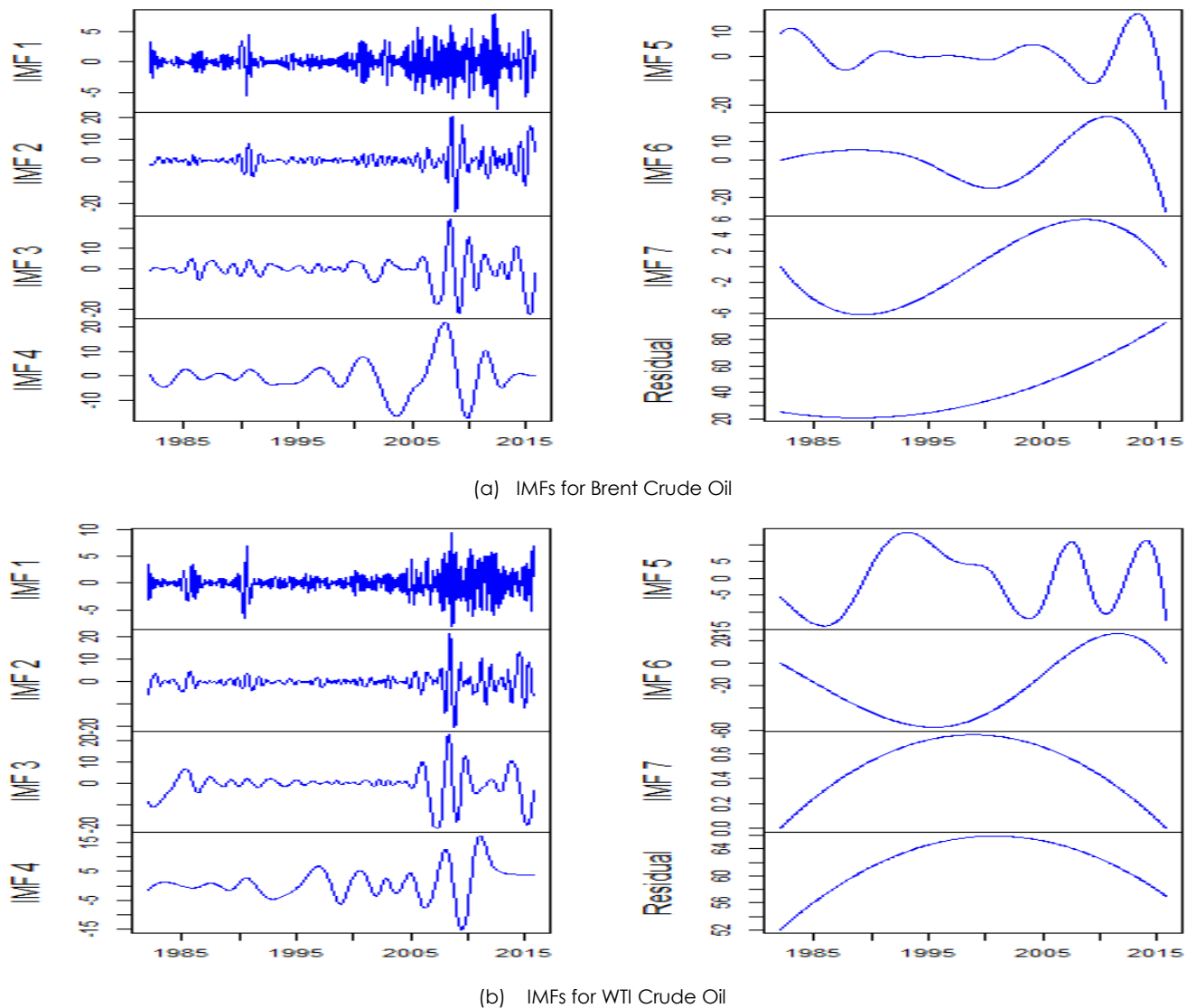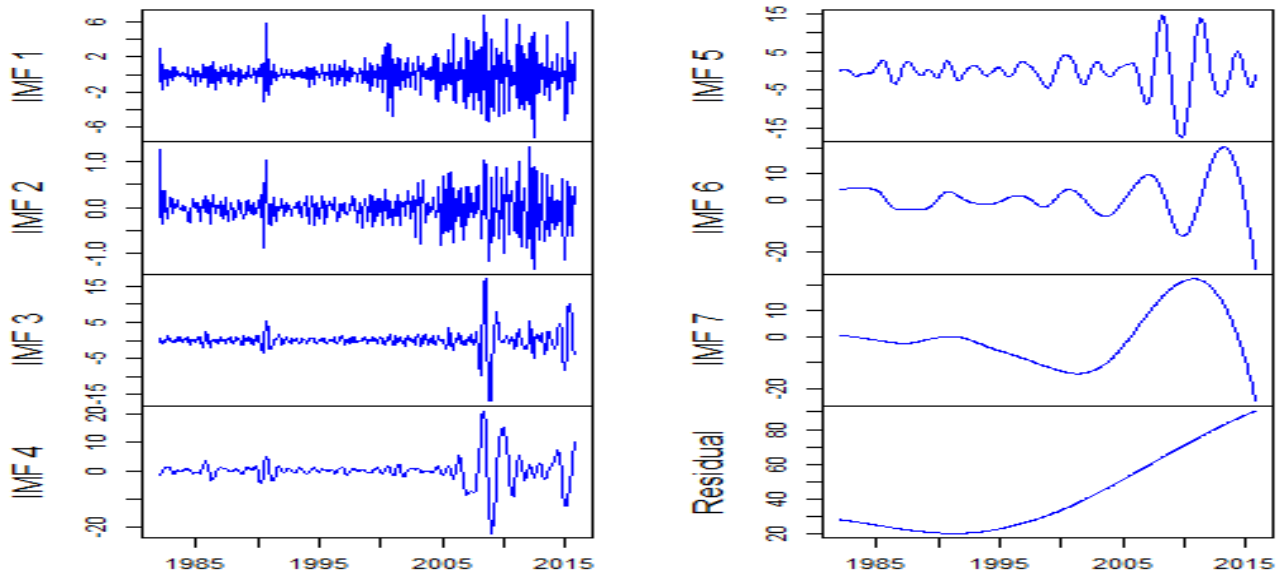

(a)   IMFs for Brent Crude Oil

(b)   IMFs for WTI Crude Oil

Figure 2 The Price Decomposition using EMD Technique

In this study, the complete intrinsic mode functions are not used which is generated by EMD/CEEMDAN technique. The purpose of the limited use of extracted IMFs is to get the better prediction. The limited use of the IMFs was decided based on MAE and RMSE while

applying ARIMA model. First, the EMD technique is used to decompose the two crude oil prices for the extraction of several independents IMFs. For WTI series the extracted total number of IMFs including residue was eight while for Brent series it was nine. Applying

ARIMA model to a different number of IMFs, the minimum MAE and RMSE was obtained by using five IMFs, where the last IMF called residue is the sum of last IMFs and residue. The graphical representation of IMFs by EMD method are shown in Figure 2.

(a) IMFs for Brent Crude Oil
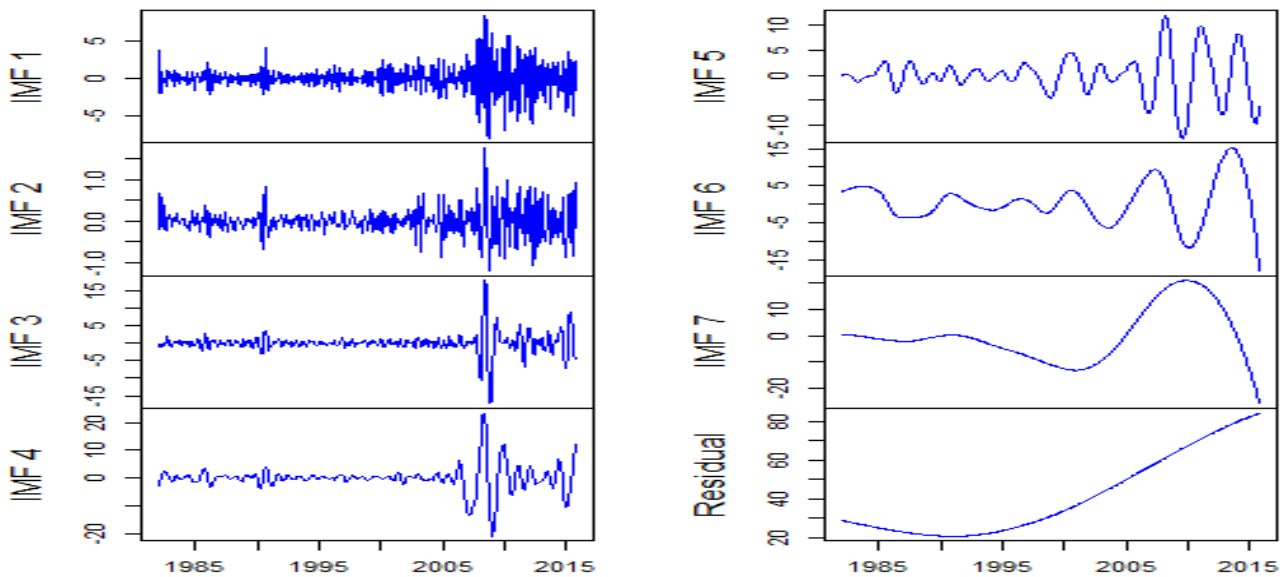
(b) IMFs for WTI Crude Oil

Figure 3 The Price Decomposition using CEEMDAN Technique

Next, the CEEMDAN technique is used to decompose the two crude oil prices series into different independent IMFs and the same number of IMFs were extracted by this method as well. For Brent series, the adaptive noise strength was taken as 0.21 while for WTI series it was equal to 0.19 with these two values the optimum mean absolute error, root mean square error and mean absolute percentage error were obtained for selected number of IMFs. Applying

ARIMA model to different numbers of IMFs extracted by CEEMDAN, the minimum MAE and RMSE was obtained by using five IMFs, where the last IMF called residue is the sum of last IMFs and residue. The graphical representation of IMFs by CEEMDAN method are shown in Figure 3. It is to be noted that only the single ARIMA and ARIMA-Kalman Filter model use the original series for prediction purposes. The rest of the ensemble models are used the extracted

decomposed IMFs for the prediction purposes for both crude oil prices series. The ARIMA model estimated parameters results are shown in Table 1 for Brent and WTI crude oil price series respectively. For the remaining four models, different orders of autoregressive and moving averages terms were used by ARIMA models for different extracted IMFs in each case. After selection of best ARIMA model for each extracted IMFs, the respective parameters were estimated in all four models and rest of the analysis performed correspondingly as discussed earlier in this section. At the end, all the IMFs and residuals results are simply added for obtaining the prediction results. Next, the evaluation of the prediction results of the two benchmark crude oil prices series are shown in Table 2 and 4 via MAE, RMSE, MAPE, and D statistic. So, it is clear from Table 3 that the CEEMDAN-ARIMA-Kalman Filter model has very promising results as compared to the other designated models listed in this study. Hence, the prediction performance of the model CEEMDAN-ARIMA-Kalman Filter is much better than other models by the output of the three-evaluation criterion MAE, RMSE, and MAPE. Concentrating on the

MAE, RMSE and MAPE indicators, the CEEMDAN-ARIMA-Kalman Filter model executes the best performance for all oil data sets, followed by the CEEMDAN-ARIMA model, EMD-ARIMA model, EMD-ARIMA-Kalman Filter, ARIMA-Kalman Filter and at the end the individual ARIMA model. Remarkably, it is interesting to note that the performance of the EMD-ARIMA model is better than the EMD-ARIMA-Kalman Filter model for both crude oil prices series. From the above results, it is also noted that the decomposition results of the IMFs of two methods that are EMD and CEEMDAN are different. Especially in the case of the models CEEMDAN-ARIMA-Kalman Filter and EMD-ARIMA-Kalman Filter, the Kalman Filter produces good forecasting results with less MAE, RMSE, and MAPE for the model CEEMDAN-ARIMA-Kalman Filter while for the model EMD-ARIMA-Kalman Filter the situation was different. From this study, we noted that the Kalman Filter is not producing good results with EMD while with CEEMDAN it produces good results. The possible reasons may be that the CEEMDAN decomposition influences the forecasting performance.

Table 2 The MAE, RMSE and MAPE comparisons of the different selected methods

Models	Brent				WTI			
	MAE	RMSE	MAPE	Rank	MAE	RMSE	MAPE	Rank
CEEMDAN-ARIMA-Kalman Filter	1.95	2.68	2.62	1	1.27	1.59	1.59	1
CEEMDAN-ARIMA	2.19	3.07	2.93	2	1.32	1.66	1.63	2
EMD-ARIMA	3.62	4.43	4.47	3	2.94	3.54	3.85	3
EMD-ARIMA-Kalman Filter	3.85	4.78	5.67	4	2.99	3.57	3.92	4
ARIMA-Kalman Filter	5.65	7.19	6.57	5	5.35	6.49	6.45	5
ARIMA	6.04	7.46	6.95	6	5.37	6.55	6.89	6

This paper also examines the percentage relative efficiency (PRE) approach used to compare different models with the basic ARIMA model per the percentage improvement in each MAE, RMSE and MAPE criterion respectively. Table 3 presented the PRE-of all models, it is clear from the Table 3 that the forecasting power of the model CEEMDAN-ARIMA-Kalman Filter is improved almost two hundred percent as compared to the single ARIMA model by the evaluation criterion MAE, RMSE, and MAPE for the

Brent crude oil series. For WTI series the results for the model CEEMDAN-ARIMA-Kalman Filter are improved three hundred percent associated with the single ARIMA model per the forecasting power of the model. The rest of the models forecasting power percentages are shown in Table 3 for both Brent and WTI crude oil prices associated with the single ARIMA model. Figure 4 also shows the percentage relative efficiency of all models with the three-evaluation criterion namely MAE, RMSE, and MAPE.

Table 3 The percentage relative efficiency comparisons of the different selected methods

Models	Brent				WTI			
	MAE	RMSE	MAPE	Rank	MAE	RMSE	MAPE	Rank
CEEMDAN-ARIMA-Kalman Filter	308.9	296.8	284.8	1	422.8	411.9	432.9	1
CEEMDAN-ARIMA	274.9	259.5	254.4	2	406.8	394.6	421.8	2
EMD-ARIMA	166.8	180.0	167.2	3	182.6	185.0	178.8	3
EMD-ARIMA-Kalman Filter	151.8	162.0	131.9	4	179.2	181.8	175.8	4
ARIMA-Kalman Filter	103.3	102.8	113.9	5	100.4	100.9	106.8	5
ARIMA	100.0	100.0	100.0	6	100.0	100.0	100.0	6

The next evaluation criterion is $D_{statistic}$ which evaluates the directional performance of the model. Since the MAE, RMSE, and MAPE do not ensure the high success rate of crude oil price movement direction which is more important for investors because the businessmen plan their activity per the market trend. So, therefore, the $D_{statistic}$ assessment is obligatory.

Table 4 presented the percentage values of $D_{statistic}$, from the Table 4 we observed that the model CEEMDAN-ARIMA-Kalman Filter perform well with high percentage among other models for both crude oil prices. Regarding the investor point of view, the $D_{statistic}$ is much important evaluation criterion than the MAE, RMSE, and MAPE for prediction purposes. According with Table 4 shows a significant difference among different models although the ranks remain same between all evaluation criterion in Table 4, the Brent crude oil $D_{statistic}$ values for single ARIMA model is 44.44%, for the ARIMA-Kalman Filter model it is 75.00%, for the EMD-ARIMA model is 82.22%, for the EMD-ARIMA-Kalman Filter model the value is 78.89%, for the CEEMDAN-ARIMA model the value is 87.78%, and for the last model which is CEEMDAN-ARIMA-Kalman Filter, the $D_{statistic}$ value is 93.33%. Except for the ARIMA model, the rest of the models are having higher percentages for $D_{statistic}$. For WTI series the pattern is the same for $D_{statistic}$ with the lowest value for ARIMA model. The main reason is that the nonlinearity, high noise, and complex dynamics are the characteristics of crude oil price series Yu andWang[38], while ARIMA is a class of linear models. The study presented in this paper is the comparison of the three different techniques the first one is the single ARIMA model, the second is the hybrid model of ARIMA and Kalman Filter while the third is the decomposition and ensemble strategy. From the outcomes of this paper, we observed that the decomposition of the series is a better strategy regarding the prediction of the series as compared to other single models.

Table 4 The percentage $D_{statistic}$ comparisons of the different selected methods

Models	Brent		WTI	
	$D_{statistic}$	Rank	$D_{statistic}$	Rank
CEEMDAN-ARIMA-Kalman Filter	93.33	1	89.29	1
CEEMDAN-ARIMA	87.78	2	85.71	2
EMD-ARIMA-Kalman Filter	78.89	4	78.57	4
EMD-ARIMA	82.22	3	82.14	3
ARIMA-Kalman Filter	75.00	5	65.45	5
ARIMA	44.44	6	50.00	6

In this paper, the hybrid model of ARIMA-Kalman Filter is used to predict the extracted IMFs from the original series. So, the performance of the model CEEMDAN-ARIMA-Kalman Filter is considerably better than other models considered in this study, namely they are single ARIMA model, ARIMA-Kalman Filter model, EMD-ARIMA model, EMD-ARIMA-Kalman Filter model and CEEMDAN-ARIMA model for the mentioned two crude oil prices with respect to forecasting power accuracy, as stately measured by MAE, RMSE, MAPE and the directional $D_{statistic}$.

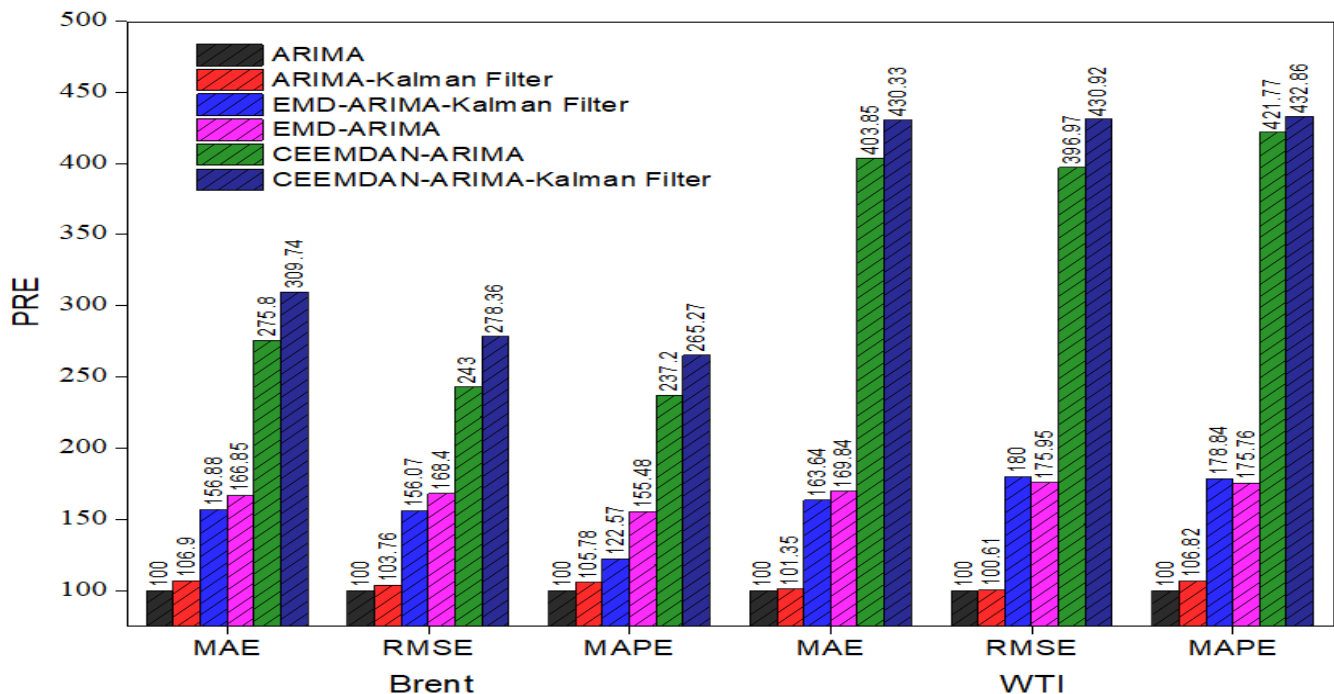**Figure 4** Comparisons of PRE of Brent and WTI series

4.0 CONCLUSION

This paper proposed a CEEMDAN-ARIMA-Kalman Filter model for forecasting the world crude oil monthly prices. In terms of experimental results, the suggested model performed well by evaluating the three descriptive measures that are MAE, RMSE, MAPE and one directional measure that is $D_{statistic}$ criterion. On every case the MAE, RMSE, and MAPE are the lowest while the $D_{statistic}$ is the highest, demonstrating that the CEEMDAN-ARIMA-Kalman Filter technique is a promising approach needs to be used for prediction of the world crude oil monthly prices in future.
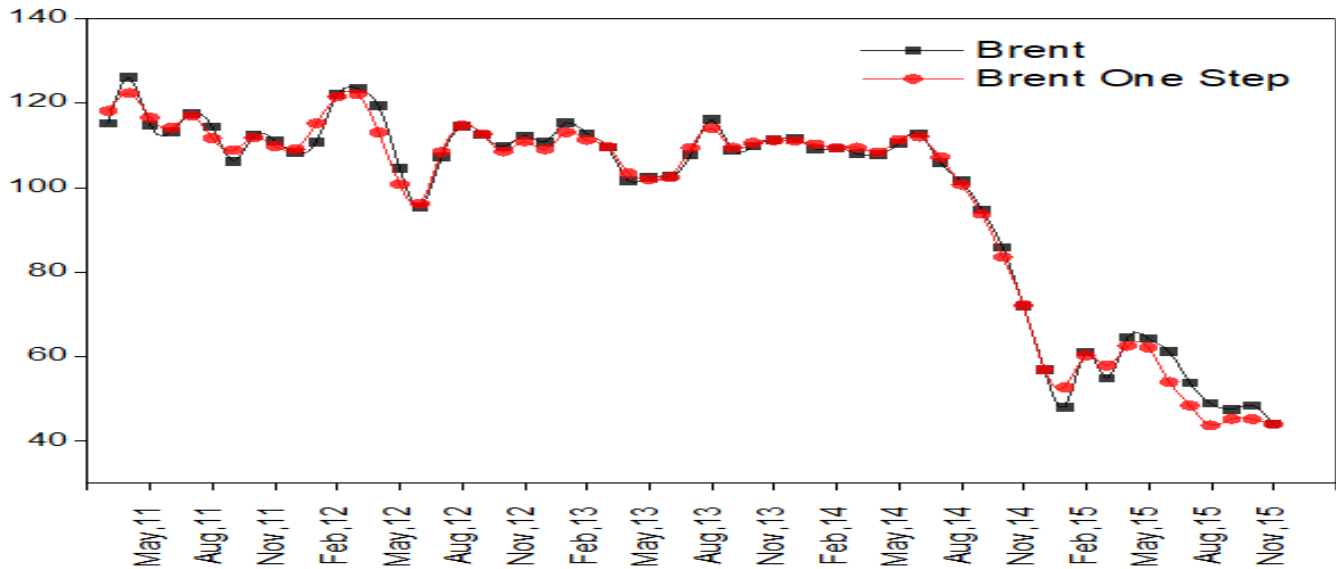
(a) One step ahead forecast for testing data of Brent crude oil prices
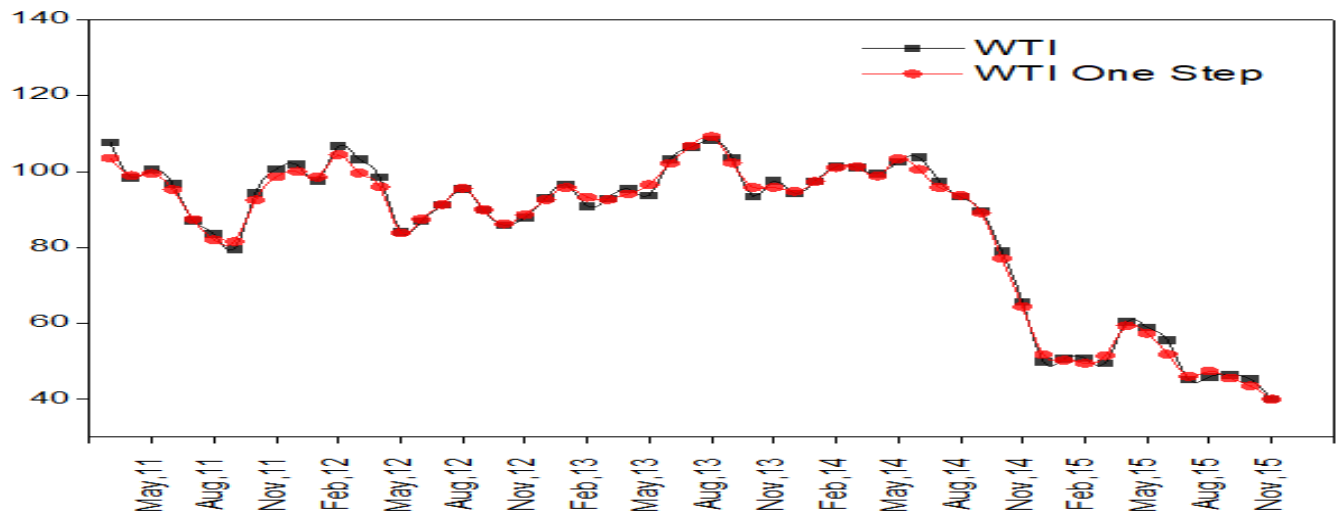
(b) One step ahead forecast for testing data of WTI crude oil prices

Figure 5 One Step Ahead Forecast for Testing Data Sets of both Crude Oil Prices

References

[1]   Watkins, G. and Plourde, A. 1994. How Volatile Are Crude Oil Prices. *OPEC Review*. 18(4): 220-245.
[2]   Alvarez-Ramirez, J., *et al*. 2003. Symmetry/Anti-Symmetry Phase Transitions in Crude Oil Markets. *Physica A: Statistical Mechanics and its Applications*. 322: 583-596.
[3]   Verleger, P. 1994. *Adjusting to Volatile Energy PricesInstitute for International Economics*. Washington.
[4]   Hagen, R. 1994. How is the International Price of a Particular Crude Determined? *OPEC Review*. 18(1): 127-135.
[5]   Stevens, P. 1995. The Determination of Oil Prices 1945–1995: A Diagrammatic Interpretation. *Energy Policy*. 23(10): 861-870.

[6]   Abosedra, S. and Baghestani, H. 2004. On the Predictive Accuracy Of Crude Oil Futures Prices. *Energy Policy*. 32(12): 1389-1393.

[7]   Wang, S. Y., L. A. Yu and K. K. Lai. 2005. Crude Oil Price Forecasting with TEI@ I Methodology. *Journal of Systematic Science. Complexity*. 18(2): 145-166.

[8]   Huntington, H. G. 1994. Oil Price Forecasting in the 1980s: What Went Wrong? *The Energy Journa*l. 1-22.

[9]   Abramson, B. and Finizza, A. 1995. Probabilistic Forecasts from Probabilistic Models: A Case Study in the Oil Market. *International Journal of Forecasting*. 11(1): 63-72.

[10]   Barone-Adesi, G., Bourgoin, F., and Giannopoulos, K. 1998. Don't Look Back. *Risk*. 11:100-104.

[11]   Morana, C. 2001. A Semiparametric Approach to Short-Term Oil Price Forecasting. *Energy Economics*. 23(3): 325-338.

[12]   Gülen, S. G. 1998. Efficiency in the Crude Oil Futures Market. *Journal of Energy Finance & Development*. 3(1): 13-21.

[13]   Mirmirani, S. and Li, H. C. 2004. A Comparison of VAR and Neural Networks with Genetic Algorithm in Forecasting Price of Oil. *Advances in Econometrics*. 19: 203-223.

[14]   Lanza, A., Manera, M., and Giovannini, M. 2005. Modeling and Forecasting Cointegrated Relationships Among Heavy Oil and Product Prices. *Energy Economics*. 27(6): 831-848.

[15]   Ye, M., Zyren, J., and Shore, J. 2002. Forecasting Crude Oil Spot Price using OECD Petroleum Inventory Levels. *International Advances in Economic Research*. 8(4): 324-333.

[16]   Ye, M., Zyren, J., and Shore, J. 2005. A Monthly Crude Oil Spot Price Forecasting Model Using Relative Inventories. *International Journal of Forecasting*. 21(3): 491-501.

[17]   Ye, M., Zyren, J., and Shore, J. 2006. Forecasting Short-run Crude Oil Price Using High-and Low-inventory Variables. *Energy Policy*. 34(17): 2736-2743.

[18]   Dées, S., *et al*. 2008. Assessing the Factors Behind Oil Price Changes. European Central Bank Working Paper. 4-36.

[19]   Liu, L. M. 1991. Dynamic Relationship Analysis of US Gasoline and Crude Oil Prices. *Journal of Forecasting*. 10(5): 521-547.

[20]   Chinn, M. D., LeBlanc, M., and Coibion, O. 2005. The Predictive Content of Energy Futures: An Update on Petroleum, Natural Gas, Heating Oil and Gasoline. National Bureau of Economic Research.

[21]   Agnolucci, P. 2009. Volatility in Crude Oil Futures: A Comparison of the Predictive Ability of GARCH and Implied Volatility Models. *Energy Economics*. 31(2): 316-321.

[22]   Ahmad, M. 2012. Modelling and Forecasting Oman Crude Oil Prices Using Box–Jenkins Techniques. *International Journal of Trade and Global Markets*. 5(1): 24-30.

[23]   Sadorsky, P. 2006. Modeling and Forecasting Petroleum Futures Volatility. *Energy Economics*. 28(4): 467-488.

[24]   Hou, A. and Suardi, S. 2012. A Nonparametric GARCH Model of Crude Oil Price Return Volatility. *Energy Economics*. 34(2): 618-626.

[25]   Ahmed, R. A. and Shabri, A. B. 2013. Fitting GARCH Models to Crude Oil Spot Price Data. *Life Science Journal*. 10(4).

[26]   Aamir, M. and Shabri, A. B. 2015. Modelling and Forecasting Monthly Crude Oil Prices of Pakistan: A Comparative Study of ARIMA, GARCH and ARIMA-GARCH Models. *Science International*. 27(3): 2365-2371.

[27]   Diebold, F. X. and Mariano, R. S. 2012. Comparing Predictive Accuracy. *Journal Of Business & Economic Statistics*. 20(1): 134-144.

[28]   Hyndman, R., *et al*. 2008. *Forecasting with Exponential Smoothing: The State Space Approach*. Springer Science & Business Media.

[29]   Harvey, A. C. 1984. A Unified View of Statistical Forecasting Procedures. *Journal of Forecasting*. 3(3): 245-275.

[30]   Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Fluids Engineering*. 82(1): 35-45.

[31]   Morrison, G. W. and Pike, D. H. 1977. Kalman Filtering Applied to Statistical Forecasting. *Management Science*. 23(7): 768-774.

[32]   Rosenberg, B. 1973. Random Coefficients Models: The Analysis of a Cross Section of Time Series by Stochastically Convergent Parameter Regression. *Annals of Economic and Social Measurement*. 2(4): 399-428.

[33]   Engle, R. F. 1979. Estimating Structural Models of Seasonality. *Seasonal Analysis of Economic Time Series*. NBER. 281-308.

[34]   Harvey, A. C. and Phillips, G. D. 1979. Maximum Likelihood Estimation of Regression Models with Autoregressive-Moving Average Disturbances. *Biometrika*. 66(1): 49-58.

[35]   Ravichandran, S. and Prajneshu, J. 2001. State Space Modeling Versus ARIMA Time-Series Modeling. *Journal of Indian Society of Agricultural Statistics*. 54(1): 43-51.

[36]   Nikolaisen Sävås, F. 2013. Forecast Comparison of Models Based on SARIMA and the Kalman Filter for Inflation. Master Thesis, Uppsala Universitet.

[37]   Aamir, M. and Shabri, A. 2016. Modelling and Forecasting Monthly Crude Oil Price of Pakistan: A Comparative Study of ARIMA, GARCH and ARIMA Kalman Model. *Advances in Industrial and Applied Mathematics: Proceedings of 23rd Malaysian National Symposium of Mathematical Sciences (SKSM23)*. AIP Publishing.

[38]   Yu, L., Wang, S., and Lai, K. K. 2008. Forecasting Crude Oil Price with an EMD-based Neural Network Ensemble Learning Paradigm. *Energy Economics*. 30(5): 2623-2635.

[39]   Tang, B., Dong, S., and Song, T. 2012. Method for Eliminating Mode Mixing of Empirical Mode Decomposition Based on the Revised Blind Source Separation. *Signal Processing*. 92(1): 248-258.

[40]   An, X., *et al*. 2012. Short-term Prediction of Wind Power Using EMD and Chaotic Theory. *Communications in Nonlinear Science and Numerical Simulation*. 17(2): 1036-1042.

[41]   Guo, Z., *et al*. 2012. Multi-step Forecasting for Wind Speed Using a Modified EMD-based Artificial Neural Network Model. *Renewable Energy*. 37(1): 241-249.

[42]   Huang, N. E. and Wu, Z. 2008. A Review on Hilbert-Huang Transform: Method and Its Applications to Geophysical Studies. *Reviews of Geophysics*. 46(2).

[43]   Schlotthauer, G., Torres, M. E., and Rufiner, H. L. 2009. A New Algorithm for Instantaneous F 0 Speech Extraction Based on Ensemble Empirical Mode Decomposition. *Signal Processing Conference, 2009 17th European*. IEEE.

[44]   Wu, Z. and Huang, N. E. 2009. Ensemble Empirical Mode Decomposition: A Noise-assisted Data Analysis Method. *Advances in Adaptive Data Analysis*. 1(01): 1-41.

[45]   Wu, Z., *et al*. 2011. On the Time-varying Trend in Global-Mean Surface Temperature. *Climate Dynamics*. 37(3-4): 759-773.

[46]   Torres, M. E., *et al*. 2011. A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise. *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE.

[47]   Antico, A., Schlotthauer, G., and Torres, M. 2014. Analysis of Hydroclimatic Variability and Trends Using a Novel Empirical Mode *Decomposition: Application to the Paraná River Basin*. *Journal of Geophysical Research: Atmospheres*. 119(3): 1218-1233.

[48]   Montgomery, D. C., Jennings, C. L., and Kulahci, M. 2015. *Introduction to Time Series Analysis and Forecasting*. John Wiley & Sons.

[49]   Box, G. E., Jenkins, G. M., and Reinsel, G. C. 2011. *Time Series Analysis: Forecasting and Control*. Vol. 734. John Wiley & Sons.

[50]   Brocklebank, J. and Dickey, D. A. 2003. *SAS for Forecasting Time Series*. Vol. 2. SAS institute.

[51]   Hamilton, J. D. 1994. *Time Series Analysis*. Vol. 2. Princeton university Press Princeton.

[52]   Saini, N. and Mittal, A. K. 2014. Forecasting Volatility in Indian Stock Market using State Space Models. *Journal of Statistical and Econometric Methods*. 3(1): 115-136.

[53]   Huang, N. E., *et al*. 1998. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proceedings of the*

Royal Society of London A: Mathematical, Physical and Engineering Sciences. The Royal Society.

[54] Huang, N. E., Shen, Z., and Long, S. R. 1999. A New View of Nonlinear Water Waves: The Hilbert Spectrum 1. *Annual Review of Fluid Mechanics*. 31(1): 417-457.

[55] Colominas, M. A., *et al.* 2012. Noise-assisted EMD Methods in Action. *Advances in Adaptive Data Analysis*. 4(04): 1250025.

[56] Yu, L., Wang, S., and Lai, K. K. 2005. A Novel Nonlinear Ensemble Forecasting Model Incorporating GLAR and ANN for Foreign Exchange Rates. *Computers & Operations Research*. 32(10): 2523-2541.

[57] Yu, L., Wang, S., and Lai, K. K. 2010. *Foreign-Exchange-Rate Forecasting with Artificial Neural Networks*. Vol. 107. Springer Science & Business Media.

[58] Petris, G. 2010. An R Package for Dynamic Linear Models. *Journal of Statistical Software*. 36(12): 1-16.

[59] Akaike, H. 1974. A New Look at the Statistical Model Identification. *Automatic Control, IEEE Transactions on*. 19(6): 716-723.